UNIVERSIDAD PABLO DE OLAVIDE, DE SEVILLA

DEPARTAMENTO DE ECONOMÍA, MÉTODOS CUANTITATIVOS E HISTORIA ECONÓMICA

Propuesta de valoración de las influencias entre Educación y Economía

Memoria presentada por Inmaculada Romano Paguillo para optar al grado de Doctora por la Universidad Pablo de Olavide, de Sevilla.

Fdo.: Inmaculada Romano Paguillo

Vº. Bº. EL DIRECTOR:

Fdo.: Eugenio M. Fedriani Martel

Profesor Titular de Universidad del Departamento de Economía, Métodos Cuantitativos e Historia Económica de la Universidad Pablo de Olavide, de Sevilla.

Sevilla, junio de 2014

Dedicado a mis seres queridos

Agradecimientos

No sé cómo puedo expresar mi más profundo y sincero agradecimiento a tantas personas que me han ayudado durante todo el desarrollo de esta tesis.

En primer lugar quiero agradecer la labor de mi director de tesis, el profesor Dr. Eugenio M. Fedriani Martel. Este trabajo se ha podido llevar a cabo gracias a la labor incondicional de Eugenio, por su paciencia, por su apoyo en todo momento, por sus indicaciones, por su motivación, por sus palabras de ánimo para seguir trabajando y por la gran ayuda que me ha prestado durante todo el proceso. Gracias a Eugenio he podido desarrollar este trabajo de investigación. Muchas gracias, Eugenio.

Mis agradecimientos también a mis compañeros del Departamento de Economía, Métodos Cuantitativos e Historia Económica, por su ayuda y colaboración, en particular a su Directora, la Dra. Flor Mª. Guerrero Casas, por su apoyo, desde el momento en que empecé a trabajar en el Departamento, por la oportunidad que me ofreció. Muchas gracias, Flor.

A todos los compañeros del Área de Métodos Cuantitativos, por su colaboración para poder llevar a cabo esta tesis y por su apoyo y las palabras de ánimo de muchos de ellos. También a Isabel Alonso y Jesús Trujillo por toda su ayuda

desinteresada.

A mis padres, que han estado en todo momento ofreciéndome su apoyo incondicional desde el día que les dije que me matriculaba en Matemáticas. Desde ese momento ellos me han ayudado y me han dado las fuerzas para afrontar distintos retos, hasta llegar a este. Para mí es muy importante que estén aquí a mi lado. Muchas gracias, Papá y Mamá.

A mi pequeña Carmen, que en los momentos en que necesitaba una pequeña sonrisa y ausentarme, ella me ha ayudado a seguir adelante, gracias a su inocencia. Gracias, Lola y Migue.

A mi compañero, amigo y pareja Jonathan, por su dedicación en todo momento. Gracias por sus palabras de ánimo y apoyo, gracias por soportar los días malos y buenos, por aguantarme y, sobre todo, por escucharme en mis ratitos de estrés. Gracias por soportarme, por quererme y darme fuerza en todo momento.

No puedo olvidar agradecer a todos mis seres queridos, como mi familia política, como a todos mis familiares y, cómo no, recordar a mis familiares que ya no están aquí presentes pero que sé que estarían muy orgullosos de mí si estuvieran a mi lado en este momento.

Mil gracias a todos, de todo corazón.

Índice general

Índice general iv

1. Introducción **1**

1.1. Economía de la Educación 1

1.2. Motivación 5

1.3. Objetivos 9

1.4. Estructura de la tesis y contribuciones 10

I METODOLOGÍA **13**

2. Introducción a las redes neuronales artificiales **15**

2.1. Redes neuronales biológicas 17

2.2. Origen de las RNA 20

2.2.1. Historia 20

2.3. Objetivos de las RNA 26

2.4. Definiciones de RNA 27
2.4.1. Revisión 27
2.4.2. Propuesta de definición de RNA 31
2.5. Propiedades 43
2.6. Tipos 45
2.6.1. Según el número de entradas o salidas 45
2.6.2. Según la estructura topológica subyacente 47
2.6.3. Según la familia de funciones neuronales 50
2.6.4. Según la regla de aprendizaje 51
2.7. Ejemplos de RNA 53
2.7.1. Asociador lineal 53
2.7.2. Perceptrón simple 54
2.7.3. Adaline 56
2.7.4. Madaline 57
2.7.5. Perceptrón multicapa (PM) 58
2.7.6. Función de base radial 60
2.7.7. Red de Hopfield 61
2.7.8. Red continua de Hopfield 62
2.7.9. Learning Vector Quantization Network 63

2.7.10. Red probalística 63

2.7.11. Red de regresión general 65

2.7.12. Brain State in a Box 65

2.7.13. Mapa autoorganizado 66

2.7.14. Máquina de Boltzmann 67

2.7.15. Adaptive Resonance Theory Network 68

2.7.16. Memoria asociativa bidireccional 69

2.7.17. Máquina de Cauchy 70

2.7.18. Mapa de Kohonen 71

2.7.19. Counter Propagation Network 72

2.7.20. Red neuronal competitiva no supervisada 73

2.8. Aplicaciones 73

2.8.1. Aplicaciones en el ámbito económico o empresarial 76

3. Propuestas de ajustes de la técnica 81

3.1. Tratamiento de bases de datos incompletas 81

3.1.1. Preliminares 81

3.1.2. Introducción al problema de clasificación con datos perdidos 82

3.1.3. Análisis de datos completos (*listwise* o *case deletion*, LD) . 85

3.1.4. Análisis de datos disponibles (*pairwise deletion*, PD) 87

3.1.5. Análisis con imputación de datos 88

3.1.6. Subsanación mediante RNA 90

3.2. Determinación de la RNA más adecuada 100

3.2.1. Implementación en Mathematica 103

3.2.2. Ejemplo de búsqueda de la RNA más adecuada 120

3.3. Reducción de parámetros y tiempo computacional 138

3.3.1. Ejemplo de simplificación de la estructura topológica 139

3.4. Propuesta de delimitación de técnicas 146

II APLICACIÓN **151**

4. Presentación del problema **153**

4.1. Análisis del problema por parte de otros autores 155

5. Datos **159**

5.1. Datos educativos 160

5.2. Características socio-económicas 177

5.3. Análisis descriptivos 187

5.4. Sobre la procedencia de los estudiantes 191

5.4.1. Descripción preliminar de los datos 192

5.4.2. Diferentes análisis de los datos 199

6. Aplicación **217**

6.1. Índice del rendimiento académico δ 218

6.1.1. Cálculo del índice del rendimiento académico δ 222

6.1.2. Análisis estadísticos 223

6.1.3. Diseño de la ruta académica óptima 227

6.2. Comprobación de resultados mediante fsQCA 241

6.2.1. Descripción de la técnica fsQCA 242

6.2.2. Resultados de la aplicación de fsQCA 249

7. Conclusiones **253**

7.1. Resultados metodológicos 254

7.2. Resultados de la aplicación 255

7.3. Otras propuestas de investigación 257

Bibliografía **259**

A. Datos ejemplos **277**

B. Datos objetivos **283**

C. Resultados parciales **293**

Lista de figuras **299**

Lista de tablas **303**

Capítulo 1

Introducción

1.1. Economía de la Educación

En la época en que vivimos, la Economía de la Educación se está convirtiendo en una de las claves para intentar transformar la sociedad de una forma fundamentada y consciente. Es una disciplina que estudia las reglas por las que se rigen la producción, la distribución y el consumo de bienes y servicios educativos; también analiza los efectos socioeconómicos de los cambios que se puedan producir en dichos bienes y servicios.

Podemos afirmar, sin temor a equivocarnos, que prácticamente todos los seres humanos contemporáneos hemos tenido una relación directa con decisiones sobre educación, pues nos hemos educado de uno u otro modo; además, la mayoría hemos influido de forma indirecta en los sistemas educativos, cuando elegimos el Gobierno según sus políticas educativas. Aunque todos nos vemos capaces, en algún momento de nuestras vidas, de valorar subjetivamente las prácticas educa-

tivas o sus resultados, muy pocos se atreven a proponer medidas objetivas de la validez de la educación o a describir las relaciones causa-efecto entre las decisiones que afectan a los sistemas educativos y los resultados educativos, culturales, sociales, económicos...

La importancia de la Economía de la Educación se entiende tanto en países desarrollados como en aquellos que no lo están. La Educación es, en cualquier caso, un elemento esencial para generar una sociedad más integrada y próspera. Sin embargo, a menudo es difícil contar con herramientas adecuadas que ayuden a entender correctamente cuáles son las reglas antes aludidas y de qué forma afectan las decisiones educativas en la educación efectiva de la sociedad o cómo afecta el sistema educativo en el desarrollo económico.

Es más, tampoco podemos olvidar la importancia de la relación inversa, por la cual la Economía permite una mejor implementación del sistema educativo deseado. A pesar de que los agentes sociales están de acuerdo en que es conveniente mejorar el sistema educativo, cada sociedad dedica un empeño distinto e, incluso, diseña medidas que parecen dirigir a resultados muy distintos. No es sencillo convencer a la clase política de cuál es la mejor forma de apoyar económicamente la Educación, como tampoco está nada claro cuál es el papel concreto (cuantificado objetivamente) que la Educación jugará en el futuro desarrollo socioeconómico.

De un modo esquemático, la Economía de la Educación se dedica a analizar varios puntos:

- El capital humano y las tasas de rendimiento de la educación;
- la educación, la formación y la inserción laboral;
- el nivel educativo y las trayectorias laborales;

- los desajustes entre la oferta y la demanda de competencias y cualificaciones;
- la financiación de la educación;
- la educación y el crecimiento económico.

En cuanto a la situación actual del sistema económico y educativo en nuestro entorno, conviene recalcar que estamos inmersos en una profunda crisis económica y de valores a nivel internacional, que también ha afectado y afecta de forma significativa a España; paralelamente, el sistema educativo español y, en particular, el andaluz han sido duramente criticados a raíz de la publicación del último informe PISA [113]. Cierto es que el sistema también fue criticado tras la publicación de los informes anteriores ([112] y [111]), pero ahora hay una mayor sensibilidad hacia estos temas por la incuestionable importancia del pilar educativo en el futuro económico y social del país. Paralelamente, se pone en duda en diferentes foros si los centros educativos están preparando a los jóvenes para las profesiones del mañana; probablemente, la respuesta más sincera tenga que ser negativa. Creemos que la situación que acabamos de describir, si cabe, proporciona un mayor interés a una investigación como la que propondremos en las siguientes páginas.

Fijémonos ahora en la cuestión más relacionada con el tratamiento de la información. Un problema común a Educación y Economía viene dado por la complejidad de la información que se maneja y la dificultad que entraña estudiar los distintos factores que afectan a los sujetos de estudio. Autores como [56], [99] y [85] han estudiado factores diversos y variados; algunos de ellos afectan a la Educación y, como consecuencia, en el sentido afirmado por [87], también a la Economía. Otros trabajos, en cambio, se fijan en factores que afectan a la Economía directamente y estudian la posible afección a la Educación (ver, por ejemplo, [90]). Algunos incluso (como los últimos trabajos del profesor Marcenaro [87])

se atreven a medir la relevancia del rendimiento educativo como motor clave del crecimiento económico. Sin embargo, es habitual en el ámbito socieconómico que los investigadores tengan unas grandes limitaciones y dificultad para esclarecer la información, por la propia complejidad inherente a los sistemas sociales. Existe un inmenso número de factores que afectan a estos sistemas sociales y, sobre todo, dichos factores están tan interrelacionados que no es nada sencillo explicar el funcionamiento de los fenómenos sociales o económicos. Ello obliga a utilizar técnicas o métodos complejos, pero a veces no es posible comprobar sus hipótesis de aplicación o, en el mejor de los casos, los resultados obtenidos no resultan fáciles de interpretar.

Es cierto que ya existe una gran diversidad de técnicas de análisis multivariante que pueden ayudar al investigador (al que las domine), pero estas técnicas son en general de naturaleza cuantitativa y tienen unos requisitos bastante rígidos, luego el primer problema que se detecta en los distintos métodos aplicables es que no permiten la incorporación de variables cualitativas o difusas, como que tampoco son utilizables en presencia de datos poco fiables o perdidos. En general, dichas técnicas no se pueden aplicar directamente, sin derivar en gran parte a la subjetividad del investigador o a la imprecisión por no validar adecuadamente las variables cualitativas y las posibles relaciones existentes entre ellas. Por otro lado, el desarrollo de metodologías *ad hoc* suele estar fuera del alcance de los investigadores expertos en el análisis de casos y en la interpretación de resultados de la aplicación de técnicas estándar.

Creemos que, a causa de todo lo anterior, en la última década se ha desarrollado con un gran auge una nueva estrategia para el análisis de datos: la utilización de la inteligencia artificial como apoyo para el investigador incapaz de asimilar información excesivamente compleja, conjuntos de datos excesivamente amplios

o variables excesivamente relacionadas. Una de las metodologías con mayor éxito de aplicación lo constituye los modelos de redes neuronales artificiales que, en esencia, son estructuras formales de carácter matemático y estadístico con la propiedad del aprendizaje, es decir, de la adquisición de conocimientos que, en la mayoría de los casos, se producen a partir de la presentación de ejemplos. Estos sistemas automáticos expertos aún presentan ciertas limitaciones, pero creemos que es posible mejorarlos (por ejemplo, como se propone en la presente memoria) para que la Informática siga siendo una herramienta útil para analizar la información y no simplemente un modo de pasar el rato o una forma de justificar nuestras opiniones.

1.2. Motivación

Como se ha insinuado ya, en la actualidad existen muchos problemas en el ámbito económico, en el educativo y, en particular, en el ámbito de la enseñanza de las Ciencias Exactas. Existe una sensación general de que estos problemas están relacionados, pero es díficil determinar dónde se podría actuar para conseguir una mejora de la situación (bien parcial o bien global). En el último informe del Programa para la Evaluación Internacional de los Estudiantes (PISA) [112], la OCDE llega a concluir que, en ámbito de las Matemáticas, los estudiantes españoles tiene un gran déficit en compresión y razonamiento. Dicho déficit dificulta la adquisición de conceptos, procedimientos y, en general, competencias esenciales para desarrollar habilidades más o menos complejas. Este es uno de los principales motivos de la necesidad de estudiar el rendimiento de los estudiantes en niveles de Educación Superior y, en particular, en asignaturas cuantitativas, claramente marcadas por la necesidad de un razonamiento matemático. Obviamente, también

sería interesante observar si el problema surge en niveles educativos de Primaria y Secundaria Obligatoria y si se transforma de algún modo con los años de formación hasta llegar a afectar a la enseñanza universitaria. Sin embargo, en este trabajo nos centraremos en la Universidad, para abrir la vía de aplicación futura de ampliarla o relacionarla con los trabajos centrados en otros niveles educativos.

De hecho, uno de los principales problemas para poder estudiar el rendimiento académico de los estudiantes ha sido tradicionalmente la escasez de datos de calidad, ya que, por ejemplo en Andalucía, hasta el año 2010 la Consejería de Educación de la Junta de Andalucía (mediante las pruebas de evaluación de diagnóstico) no se dedicó a proponer un sistema homogéneo de evaluación en toda la Comunidad en Primaria y Secundaria así como a integrarlo en el sistema de información SÉNECA y en los resultados de la Encuesta Social de Andalucía 2010 (Educación y Hogares en Andalucía, realizada por el Instituto de Estadística y Cartografía de Andalucía). Desde entonces se han podido estudiar distintos aspectos educativos y, con ellos, el rendimiento de los estudiantes que se intenta mejorar (a partir de las conclusiones obtenidas), pero todo ello, hasta la fecha, se ha centrado en Primaria y Secundaria.

En ciertas áreas parece evidente que los estudiantes españoles tienen grandes dificultades formativas desde niños y que dichas dificultades no se están resolviendo con las distintas reformas educativas que viene sufriendo el sistema. Esto nos hace pensar que, o bien las reformas no están alineadas con los diagnósticos y objetivos o bien los diagnósticos no están correctamente realizados.

Centrándonos en niveles educativos universitarios, los estudiantes ya vienen afectados por las dificultades anteriormente señaladas (así como por las distintas reformas educativas padecidas). En particular, la última reforma educativa uni-

versitaria surge del "Plan Bolonia" y, aunque lleva implementándose desde hace solo unos años, no parece que esté consiguiendo una mejora significativa del rendimiento académico. En concreto, en la Facultad de Ciencias Empresariales de la Universidad Pablo de Olavide, de Sevilla, la reforma se comenzó a aplicar en el curso académico 2009-2010, aunque ya se habían liderado a nivel nacional diferentes experiencias piloto en los cursos anteriores. Creemos conveniente destacar que los estudios de Grado de dicha Facultad de Ciencias Empresariales pueden ser considerados clave para evaluar los frutos de estas recientes reformas, pues las titulaciones impartidas en la Facultad (especialmente claro es el caso del Grado en Administración y Dirección de Empresas) tienen una relación estrechísima con el mercado laboral y, siendo valoradas por los especialistas como titulaciones que "siempre tendrán futuro", pueden resultar representativas del nivel de adaptación de los estudios universitarios a una función profesionalizante que se demanda con Bolonia con una fuerza creciente. Por otra parte, nuestra intuición nos dice que los alumnos de Grados como el de Análisis Económico o el de Finanzas y Contabilidad tendrán también una importante influencia en la transformación económica de la sociedad y, por tanto, son candidatos idóneos para valorar (cuando se incorporen en el mercado laboral) qué influencia produce la Educación Superior en la Economía.

Comentemos ahora el origen de este trabajo de investigación. Aproximadamente desde 2010, deseábamos realizar un estudio más exhaustivo del nivel de los estudiantes universitarios que deben enfrentarse a asignaturas de las ramas matemática y estadística en titulaciones íntimamente relacionadas con el mundo de la empresa. Se trataba de averiguar si los resultados obtenidos por estos estudiantes vienen afectados por los distintos factores que se ha probado que influyen a los niveles educativos inferiores. Además, se pretendía valorar si la formación previa

o los niveles socioeconómicos ambientales tenían una influencia significativa en dicho rendimiento.

Sin embargo, este deseo se topaba con diversas dificultades: por una parte, se intuía una escased de datos fiables y útiles; por otro lado, se sospechaba sobre la inexistencia de una metodología totalmente idónea. En vista de esto, tras revisar los intentos similares producidos en otros lugares, se decidió reunir un conjunto de datos apropiado y, al mismo tiempo, desarrollar una técnica ajustada a lo necesario.

Por los motivos anteriormente reseñados, se decidió partir (al menos en este primer intento, abierto a futuras ampliaciones) de los datos aportados por los docentes de las asignaturas obligatorias de contenido cuantitativo en los grados de la Facultad de Ciencias Empresariales de la Universidad Pablo de Olavide, de Sevilla. En cuanto a la técnica, se decidió intentar aprovechar la flexibilidad y potencia de las redes neuronales artificiales, aunque ello supusiera desarrollar algunas mejoras técnicas para permitir obtener resultados cercanos a lo perseguido.

Respecto a las redes neuronales, creemos conveniente comentar que en los últimos años se han desarrollado distintos métodos matemáticos en el ámbito de la Economía y la Empresa para realizar clasificaciones, aproximaciones, estimaciones y simulaciones, pero los relacionados con la inteligencia artificial tienen un atractivo especial. La teoría de las redes neuronales artificiales ha crecido en posibilidades y en aplicaciones, sin embargo, cuando se utilizan para valorar fenómenos complejos, como los referentes al análisis económico o a la validez de los sistemas educativos, suelen requerir de adaptaciones que permitan abordar con éxito este tipo de situaciones sin caer en los mismos inconvenientes que los presentes en las técnicas tradicionales.

La raíz del problema más teórico que queremos resolver en esta tesis se encuentra en que las redes neuronales se suelen aplicar tal cual, mimetizando lo que otros autores han hecho antes con problemas similares y limitando, con ello, sus posibilidades. Consideramos que la causa de esta circunstancia es que el experto en el problema práctico rara vez pueda acceder a los fundamentos de la teoría aplicada, pues ni siquiera existe una terminología unificada ni una teoría coherente, consistente universalmente aceptada y de fácil acceso a la que recurrir. Por otro lado, los expertos en redes neuronales suelen dedicar sus esfuerzos a resolver problemas con un rendimiento más inmediato (que lo que sería la mejora de un sistema educativo, por ejemplo), a menudo patrocinados por empresas a las que no interesa la difusión de los logros computacionales. Esto hace que las publicaciones sobre redes neuronales sean frecuentemente oscuras y se limiten a una exposición somera del problema y de los resultados tras el entrenamiento de una red que actúa, a ojos del lector, como una caja negra imposible de desentrañar.

Por todo lo anterior, llegamos a desarrollar esta tesis como respuesta a una situación y como posible principio de la solución de diversos problemas detectados hasta este momento.

1.3. Objetivos

Los objetivos generales de esta tesis son dos, principalmente. Por una parte, se trata de desarrollar una nueva metodología con la ayuda de las redes neuronales artificiales (o, si se quiere, adaptar las redes a problemas específicos de los conjuntos de datos con una frecuencia de aparición considerable); para ello, se trata de describir la metodología estructuradamente, se sugieren diferentes mejoras en las redes neuronales habituales y se plantean diferentes vías de investigación futura.

Por otra parte, se intenta analizar el problema educativo desde una perspectiva cuantitativa y sin perder de vista las claves económicas; para llevar a cabo este segundo objetivo, se han utilizado las distintas mejoras presentadas en la primera parte de la memoria y se han utilizado los datos a los que se aludía en la sección anterior.

El primer objetivo ha implicado el estudio de un volumen considerable de información científica (bastante inconexa, por cierto, por las propias características del área de conocimiento implicada) así como la redacción de una especie de manual sobre aquello en lo que se ha aportado algo de orden. También se han buscado algunos ejemplos para facilitar a los futuros investigadores el acercamiento a la metodología propuesta.

Por su parte, el segundo objetivo ha conllevado la definición de un nuevo indicador del rendimiento académico de los estudiantes y se ha sugerido la implementación de la "ruta más eficiente para el estudiante", que creemos que debería se tenida en cuenta tanto a nivel individual (del propio estudiante) como de la Universidad y, por supuesto, de la sociedad en la que se engloba. En particular, la situación se estudia en el contexto actual de Andalucía, pero su repercusión podría considerarse en toda nuestra sociedad y en la economía del país.

1.4. Estructura de la tesis y contribuciones

Esta memoria está dividida en dos partes. La primera se utiliza para desarrollar la metodología estudiada (resumen de lo ya conocido, propuesta de estructura y terminología, etc.) y las diferentes mejoradas propuestas. La segunda parte se dedica a la aplicación práctica de lo comentado en la primera sobre los datos

recopilados con el fin de analizar la Universiad y la Economía andaluzas.

A su vez, la primera parte de la tesis se desarrolla en dos capítulos: el primero presenta una introducción a las redes neuronales artificiales, realizando una primera revisión de la teoría existente y concluyendo con la propuesta de una definición casi-formal del concepto de red. En el capítulo siguiente, se desarrollan las distintas mejoras propuestas; en particular, destacan el tratamiento de bases de datos incompletos, la determinación de la red más adecuada para un conjunto de datos determinado y la reducción de parámetros (y del consiguiente tiempo computacional) en casos con variables explicativas no independientes.

La segunda parte está compuesta por tres capítulos. En el primero se presenta el problema a estudiar y se realiza una revisión de problemas parecidos que ya han sido afrontados por otros investigadores y en otros contextos. El segundo capitulo de esta segunda parte se describen los datos recopilados y utilizados en esta tesis; además, se presentan los primeros análisis descriptivos de los mismos. En el siguiente capítulo se proponen diferentes soluciones parciales del problema que se deseaba resolver inicialmente. El último capítulo de la segunda parte está compuesto por las conclusiones que se han deducido y de las ideas de investigación que se pretenden realizar en el futuro próximo.

La memoria concluye con las correspondientes referencias bibliográficas, un par de anexos de datos y cálculos y los correspondientes índices de figuras y tablas.

Parte I

METODOLOGÍA

Capítulo 2

Introducción a las redes neuronales artificiales

En este capítulo se trata de presentar un concepto que destaca tanto por sus aplicaciones prácticas como por la multiplicidad de aproximaciones que ha sufrido en su corta historia: el de red neuronal artificial (RNA). Inicialmente se comentará su origen, ejemplo de biomimetismo. Después se recorrerá la historia de su desarrollo. Posteriormente se tratará de aunar las características y los tipos de RNA para proponer una definición matemática ligeramente más formal que las utilizadas hasta el momento por los diferentes investigadores. En suma, se pretende recopilar y ordenar la información existente sobre las RNA, a fin de proporcionar en lo posible una base sólida y práctica a cualquiera que desee introducirse en el estudio de esta interesante área de conocimiento.

No obstante, ha de tenerse en cuenta que las RNA constituyen un campo de muy amplio y que actualmente está evolucionando a un ritmo vertiginoso. Por eso, como en el resto de esta memoria, no será posible incluir todo el material relevante,

sino solo lo necesario para comprender el ámbito de estudio y las mejoras que se propondrán para el futuro. En cualquier caso, consideramos que el contenido que se presenta puede ser útil para resolver problemas que hasta ahora no se han podido abordar con la suficiente ayuda computacional.

De hecho, el estudio de las RNA se puede considerar dentro de los límites de la inteligencia artificial (IA). Y la IA está compuesta, a su vez, por varias áreas de investigación diferenciadas, como son: los lenguajes naturales (PLN, o NLP, por las siglas en inglés de *natural language processing*), los sistemas expertos y de conocimiento (*expert systems*), la robótica, los algoritmos genéticos (*genetic algorithms*), los sistemas borrosos o difusos (*fuzzy systems*) y las RNA (*artificial neural networks*). En [26], por ejemplo, se pueden ver algunos trabajos diferenciados por área.

En cuanto a la definición de IA, es útil recurrir al Diccionario de la Real Academia Española para obtener una primera aproximación al concepto. Así, "inteligencia" es: la capacidad de entender o comprender; la capacidad de resolver problemas; el conocimiento, comprensión o acto de entender. Y el adjetivo "artificial" puede referirse a algo producido por el ingenio humano. Finalmente, define la "inteligencia artificial" como el desarrollo y utilización de ordenadores con los que se intenta reproducir los procesos de inteligencia humana. Es más, según Winston [148], uno de los primeros investigadores en IA, se debe definir como la disciplina científica y técnica que se ocupa del estudio de las ideas que permiten ser inteligentes a los ordenadores. Parece, por tanto, que la clave está en que el ser humano intenta reproducir los procesos de su propia inteligencia en otros entes, no naturales. Veamos a continuación cómo las redes neuronales biológicas han servido de inspiración para el desarrollo de las RNA.

2.1. Redes neuronales biológicas

Como era de esperar, existe una relación intrínseca entre las redes neuronales biológicas y las RNA, debido a que estas últimas se desarrollaron a partir del estudio de las primeras [92]. Por eso, creemos conveniente resumir someramente las características de los sistemas neuronales biológicos. Estos sistemas están formados por células vivas, cada una de ellas se llama *neurona* o *célula nerviosa* y está integrada por un *cuerpo celular* o *soma* del que parte el *axón* y un denso árbol de ramificaciones más cortas, compuesto por *dendritas*. Según la RAE, una neurona es una célula nerviosa, que generalmente consta de un cuerpo de forma variable y provisto de diversas prolongaciones, una de las cuales, de aspecto filiforme y más larga que las demás, es el axón o neurita. Por lo general, existen múltiples conexiones entre distintas neuronas y cada una de estas conexiones se denomina *sinapsis* o *contacto*: se producen entre el axón de una neurona y las dendritas de otras.

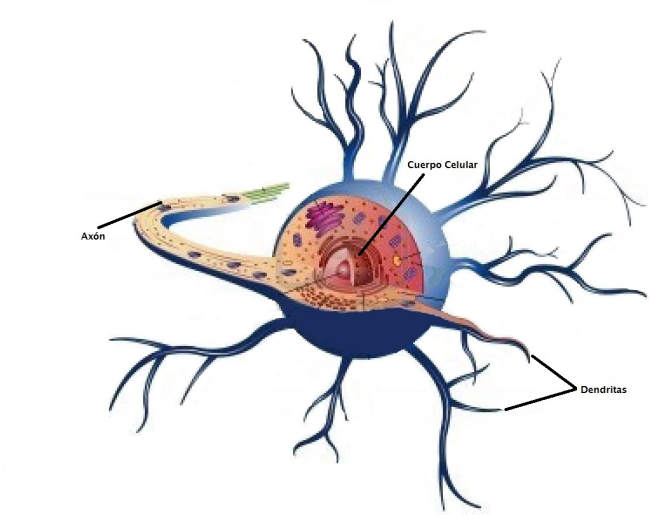


Figura 2.1: Neurona biológica, modificada de `http://es.123rf.com/`

Dentro del sistema nervioso humano, destaca por su funcionalidad el cerebro, órgano paradigmático de la inteligencia. En el origen de las RNA, se trataba de copiar la fisiología del cerebro porque es muy eficiente a la hora de reconocer patrones. Algunas de otras características relevantes del cerebro humano son las siguientes:

1. El procesamiento de información biológico es robusto y tolerante a fallos. Nuestro cerebro tiene el mayor número de neuronas que poseemos en nuestra vida en el momento del nacimiento; diariamente hay neuronas que se van "apagando", pero este hecho no hace que el cerebro deje de funcionar hasta que llegue a un grado de deterioro grave.

2. Los procesadores de información son flexibles: se reajustan por sí solos a la hora de reaccionar ante algún cambio.

3. Es capaz de trabajar con falta de información.

4. Las neuronas están posicionadas de forma paralela, son pequeñas, compactas... hasta se podría decir que se distribuyen siguiendo una estructura fractal.

5. Las neuronas son numerosísimas, pero individualmente necesitan poca cantidad de energía para funcionar.

Es obvio que cualquier informático desearía para su computador unas características parecidas a las listadas anteriormente. En la Tabla 2.1 se presentan las principales semejanzas y diferencias entre el cerebro humano y los ordenadores actuales.

Tabla 2.1: Relación entre el cerebro humano y el ordenador

Características	Cerebro humano	Ordenador
Velocidad de proceso	entre 10^{-3} y 10^{-2} s	entre 10^{-9} y 10^{-8} s
Estilo de procesamiento	en paralelo	en serie
Número de procesadores	entre 10^{11} y 10^{14}	pocos
Conexiones	10000 por procesador	pocas
Almacenamiento	distribuido	en direcciones fijas
Tolerancia a fallos	amplia	poca o nula
Tipo de control del proceso	auto-organizado	centralizado
Consumo de energía operac./s	10^{16} Julios	10^{6} Julios

Fuente: elaboración propia a partir de [27]

Como ya se ha insinuado, los ordenadores copian el funcionamiento del cerebro. Sin embargo, obviamente, existen diferencias estructurales y funcionales significativas entre cualquier cerebro humano y los ordenadores: Los ordenadores presentan habitualmente una arquitectura "de tipo Von Neumann", basada en un microprocesador muy rápido capaz de ejecutar en serie instrucciones complejas de forma fiable, mientras que el cerebro humano está compuesto por millones de procesadores elementales o neuronas, interconectadas entre sí formando redes; además, las neuronas biológicas no necesitan ser programadas, sino que aprenden a partir de los estímulos que reciben del entrenamiento y operan siguiendo un esquema masivamente paralelo, distinto al procesamiento en serie típico de los ordenadores convencionales (para más información a este respecto, se puede consultar [27]).

Las RNA también copian el funcionamiento del cerebro, aunque no tanto a nivel de *hardware* sino a través del *software* o programación. Eso no representa

un inconveniente grave, pues las RNA normalmente no son evaluadas por su estructura sino por cómo funcionan. A continuación trataremos de resumir cómo han ido evolucionando las RNA según los retos que han ido afrontándose con su ayuda.

2.2. Origen de las RNA

Puede decirse que el interés por el desarrollo de las RNA nació en el año 1943, prácticamente con los inicios de la Informática. Aunque debe tenerse en cuenta que los primeros estudios sobre el comportamiento del cerebro humano y el proceso del pensamiento (que no olvidemos que las RNA intentan replicar) se deben a filósofos griegos como Aristóteles y Platón. Posteriormente, diferentes filósofos del siglo XVII retomaron dichas ideas y Alan Turing, ya en el año 1936, fue el primer biólogo que trató de estudiar el cerebro humano desde el punto de vista computacional. Por otro lado, el primer estudio (médico) donde se demostró que el sistema nervioso está compuesto por una red de neuronas conectadas entre sí se debe a Santiago Ramón y Cajal, en 1888. Los primeros trabajos específicos sobre RNA surgieron lógicamente más tarde, como herederos de los autores anteriormente comentados: en 1943 el neurofisiólogo Warren McCullon y el matemático Walter Pitts implementaron los primeros modelos abstractos de una RNA [92].

2.2.1. Historia

Desde el año 1943 se han sucedido diferentes etapas en el desarrollo de las RNA. Así, en los años 50 del siglo XX se acrecentó el interés por las RNA, pero desde el año 1960 al 1980 se produjo una crisis de confianza entre los distintos

potenciales usuarios de dicha tecnología. En 1980 se produjo un renacimiento de las RNA y a partir de 1990 se han diseñado innumerables aplicaciones que las utilizan como única herramienta o en asociación con otras, como se puede ver en [136]. Desde esa fecha, se han desarrollado diferentes modelos que se apoyan en las RNA y que las superan en algún sentido; un ejemplo de esto son las *support vector machines* (SVMs, o máquinas de soporte vectorial o incluso máquinas de vectores de soporte), utilizadas para clasificación o regresión. En cualquier caso, las RNA (sus aplicaciones, sus características, las ideas de mejora, etc.) siguen dando que hablar entre los investigadores (informáticos, estadísticos, matemáticos, etc.); incluso, el incremento del número de referencias ha ido aumentando considerablemente con los años. Actualmente es prácticamente imposible seguir la pista de todas las mejoras que se están desarrollando para las RNA así como de todos los problemas que están ayudando a resolver. Tampoco es fácil imaginar cuáles serán los siguientes campos donde se aplicarán.

A modo de resumen histórico, a continuación tratamos de representar en una línea de tiempo los acontecimientos más destacados relacionados con las RNA hasta comienzos de la década de los 90 del siglo XX. Llama poderosamente la atención cómo varios investigadores pueden ser responsables de un mismo progreso en el campo de las RNA; esto se debe fundamentalmente a que es una teoría que está evolucionando muy deprisa y a menudo a espaldas de la comunidad científica (a veces no se encuentra cabida en las revistas tradicionales para una disciplina tan novedosa y otras veces no interesa publicar algo que permite a sus descubridores una cierta "ventaja competitiva" en cuanto al desarrollo de lucrativas aplicaciones).

- El estudio de las RNA comienza en el año 384 a.C.: Aristóteles analiza por primera vez el comportamiento del cerebro y el proceso de pensamiento.

- Del año 427 al 347 a.C. Platón sigue prestando atención a los principios filosóficos sobre el pensamiento.

- Entre los años 1596 y 1650, Descartes y otros filósofos empiristas retoman los razonamientos de Aristóteles, llegando a analizar el comportamiento del cerebro a través del pensamiento.

- En 1911 Santiago Ramón y Cajal postula que la dirección de trasmisión de la información en el cuerpo humano es determinada por la naturaleza en forma de red de células nerviosas; sus conocimientos sobre la naturaleza eléctrica de los impulsos nerviosos y su velocidad de transmisión son publicados en [117].

- En 1913 aparece el primer dispositivo hidráulico a partir de la implementación de una RNA por Russell, aunque los primeros estudios no llegan hasta algo más tarde.[128]

- En 1936 el matemático y filósofo Alan Turing estudia el cerebro humano como una forma de ver el mundo de la computación: somete a debate si las máquinas pueden o no pensar [6]; en este sentido, realiza el experimento denominado *prueba de Turing*.

- En 1943 Warren McCulloch y Walter Pitts publican su trabajo titulado *A logical calculus of the ideas immanent in nervous activity* [92]: modelan una red neuronal simple mediante circuitos eléctricos.

- En 1949 Donald Hebb desarrolla la primera ley de aprendizaje neuronal (la *regla de Hebb* o de *aprendizaje hebbiano* o *teoría de la asamblea celular*), que supone que, si dos neuronas que están interconectadas entre sí se activan al mismo tiempo, esto indica que existe un incremento en la fuerza sináptica.

La forma de corrección que emplea esta regla es incrementar la magnitud de los pesos si ambas neuronas están inactivas al mismo tiempo (para más información, se puede consultar [66]).

- En 1954 el profesor Minsky desarrolla la primera neurocomputadora [97].
- En 1959 el psicólogo Frank Rosenblatt crea el modelo de tres capas llamado *perceptrón* [123]. Como ejemplo de sus primeras aplicaciones, realiza el patrón del alfabeto con el perceptrón.
- En ese mismo año 1959 Bernard Widrow y Marcial Hoff crean dos tipos de modelos: el modelo ADALINE (ADAptative LINear Elements) y el modelo MADELINE (Multiple ADALINE). Estos dos modelos son los primeros en aplicarse a problemas reales. En particular, se utilizan para la eliminación del eco en llamadas telefónicas [147]. Otro de los adelantos producidos en este mismo año es la aparición de una nueva regla de aprendizaje denominada como de *Widrow-Hoff*.
- En 1967 Stephen Grossberg crea un nuevo tipo de "red avalancha", que consiste en tratar elementos discretos con actividades que varían con el tiempo y que satisfacen ecuaciones diferenciales continuas; el reconocimiento continuo del habla y el aprendizaje del movimiento de los brazos de un robot son algunas de sus aplicaciones. El mismo autor había publicado un estudio sobre los mecanismos de la percepción y la memoria en [63].
- En 1968 Stephen Grossberg, en compañía de Gail Carpenter, desarrolla una nueva adaptación a la teoría de ART (Adaptative Resonance Theory)[62].
- En 1969 Marvin Misnsky y Seymour Papert redactan el libro [98] sobre el perceptrón; en dicho libro, se critican los modelos matemáticos propues-

tos por Rosenblat, el creador del perceptrón, demostrándose importantes limitaciones teóricas en el aprendizaje del perceptrón.

- Entre 1969 y 1977 James Anderson desarrolla una extensión del asociador lineal, en concreto del modelo lineal *Brain State in a Box* (BSB) [5].

- En 1970 Teuvo Kohonen y Jim Anderson obtienen, de forma independiente, un modelo similar al propuesto por James Anderson: el modelo LVQ y los mapas auto-organizados [82].

- En el verano de 1979 James Anderson junto con Hinton organizan el primer encuentro de neo-conexionistas.

- En 1980 Kunihiko Fukushima desarrolla el Neocognitrón, un modelo neuronal utilizado para el reconocimiento de patrones visuales [55].

- En 1982 John Hopfield modeliza la *red Hopfield*, que reconstruye patrones y los optimiza; se trata de la primera red de tipo dinámica; además, presenta una variación del asociador lineal basado en la función de energía de Lyapunov para ecuaciones no lineales [72].

- En ese mismo año 1982 se celebra por primera vez el congreso *US-Japan Joint Conference Cooperative Competitive Neural Networks*.

- En 1983 Cohen y Grossber desarrollan el principio de la memoria direccional [20].

- En 1985 se lleva a cabo la primera reunión anual *Neuronal Networks for Computing*, en el Instituto Americano de Física.

- Un año después Rumelhart, Hunton y Williams [127] descubren el algoritmo *back-propagation*, que había sido desarrollado independientemente en el año

1974 por Paul Werbos.

- También de forma independiente, Werbos en 1974, Parker en 1985 y Lecunn en 1987, generalizan la regla de Delta, una regla de aprendizaje supervisado.
- Simultáneamente a los hitos anteriores, Rumelhart y Mc Clelland escriben el libro titulado *Parallel distributed processing: Explorations in the microstructure of cognition.*
- A partir de 1987 se celebran las reuniones anuales *Neural Information Processing Systems(NIPS)* e *International Neural Network Society*(INNS).
- En 1988 se constituye el congreso de formación internacional *Joint Conference on Neural Networks* (IJCNN).
- En 1991 se organiza por primera vez el congreso *International Conference on Artificial Neural Networks* (ICANN).
- En 1992 el investigador Halbert White [145] vincula la tecnología neuronal con la teoría del aprendizaje y de la aproximación; su interpretación econométrica es desarrollada dos años más tarde por los investigadores Kuan y White [146].

En la actualidad existe un gran número de asociaciones españolas, europeas e internacionales que fomentan la investigación en RNA y proporcionan información detallada sobre los avances en dicho campo. Algunas de estas asociaciones son: Asociación Española para la Inteligencia Artificial (AEPIA, `http://www.aepia.org/aepia/`), European Coordinating Committee for Artificial Intelligence (ECCAI, `http://eccai.org`), European Neural Network Society (ENNS, `http://www.ida.his.se/ida/enns/`), International Neural Network Society (INNS, `http:`

//www.inns.org/), Computational Intelligence Society (IEEE, http://cis.ieee.org/), etc.

2.3. Objetivos de las RNA

El objetivo fundamental de una RNA es generar un modelo que imite el comportamiento de un sistema neuronal biológico, en general, que sea capaz de solucionar problemas difíciles de resolver con técnicas convencionales. Se puede decir que ambos tipos de sistemas se caracterizan por el aprendizaje a través de la experiencia y por la extracción de conocimiento genérico a partir de un conjunto de datos.

Prácticamente no hay límites en los ámbitos de aplicación de las RNA, aunque se suele considerar que los problemas más adecuados son los de cuatro tipos: de aproximación, de estimación, de clasificación y de simulación. En particular, las RNA se puede utilizar para resolver problemas de:

- reconocimiento de patrones;
- compresión de información;
- reducción de la dimensionalidad;
- agrupamiento;
- clasificación;
- visualización;
- problemas de minería de datos.

Por su frecuencia de aparición, podemos decir que los anteriores son los más demandados entre las numerosas situaciones existentes. En cualquier caso, las RNA pueden ser consideradas cuando se den cualquiera de las condiciones siguientes:

- El número de variables o la diversidad de los datos es muy grande.
- Las relaciones entre las variables son entendidas, si acaso, vagamente.
- Las relaciones entre las variables son difíciles de describirse adecuadamente mediante los métodos convencionales.
- Las variables (o capturas) presentan semejanzas dentro de un conjunto de patrones, tal y como sucede en aplicaciones de procesamiento de señales, de control, de reconocimiento de patrones, de producción y reconocimiento del habla, en los negocios, en Medicina, etc.

2.4. Definiciones de RNA

2.4.1. Revisión

Parece ser que la primera definición de RNA fue propuesta por los investigadores Warren McCullon y Walter Pitts en el trabajo *A logical calculus of the ideas immanent in nervous activity* [92]. Se basaron en tres fundamentos: conocimientos sobre fisiología básica y funcionamiento de las neuronas en el cerebro; análisis formal de la lógica proposicional de Russell; y la teoría de la computación de Turing. A continuación se recogen varias definiciones para el concepto de RNA, proporcionadas por distintos autores y que han alcanzado considerable relevancia entre los investigadores.

Definición 2.4.1 (Kohonen, 1988). *Las RNA son redes interconectadas masivamente en paralelo de elementos simples y con organización jerárquica, las cuales intentan interactuar con los objetos del mundo real del mismo modo como lo hace el sistema nervioso biológico.*

Definición 2.4.2 (Hecht-Nielse, 1988). *Una RNA es un sistema de computación constituido por un gran número de elementos simples de procesamiento muy interconectados, que procesan información por medio de su estado dinámico como respuesta a entradas externas.*

Definición 2.4.3 (Reguero, 1995). *Se define RNA como un sistema de procesamiento de información compuesto por un gran número de elementos de procesamiento, profundamente conectados entre sí a través de canales de comunicación.*

Definición 2.4.4 (Hilera y Martínez, 1995 [68]). *Una RNA se define como una nueva forma de computación, inspirada en modelos biológicos. Es una red en la que existen elementos procesadores de información de cuyas interacciones locales depende el comportamiento del conjunto del sistema.*

Definición 2.4.5. *Una RNA es un modelo matemático compuesto por un gran número de elementos procesales organizados en niveles.*

Definición 2.4.6. *Las RNA pueden considerarse modelos de cálculo caracterizados por algoritmos muy eficientes que operan de forma masivamente paralela y permiten desarrollar tareas cognitivas como el aprendizaje de patrones para la clasificación o la optimización.*

Una RNA también se puede definir haciendo uso del concepto de grafo, como objeto integrado por un conjunto de nodos (o vértices) y de conexiones (aristas o arcos) entre los mismos. Para ello, son convenientes al menos dos definiciones de conceptos previos:

Definición 2.4.7. *Un **grafo dirigido** es aquel en el que cada una de las conexiones tiene asignada un sentido.*

Definición 2.4.8. *Un **grafo no dirigido** es aquel en el que las conexiones entre los vértices son todas bidireccionales o, simplemente, no tienen asignadas un sentido.*

Utilizando las dos definiciones anteriores, es posible formalizar la definición de RNA mediante el uso de Teoría de Grafos:

Definición 2.4.9. *Una RNA es un objeto o grafo integrado por un conjunto de vértices y de conexiones entre los mismos denominándose dirigido cuando todas las conexiones tienen asignadas un sentido y no dirigido cuando tales conexiones son bidireccionales.*

Definición 2.4.10 (Müller y Reinhardt, 1990)**.** *Una RNA se puede definir como un grafo dirigido con las siguientes restricciones o características:*

- *Los nodos se llaman elementos de proceso (EP).*
- *Los enlaces se llaman conexiones y funcionan como caminos unidireccionales instantáneos.*
- *Cada EP puede tener cualquier número de conexiones.*
- *Todas las conexiones que salgan de un EP deben tener la misma señal.*
- *Los EP pueden tener memoria local.*
- *Cada EP posee una función de transferencia que, en función de las entradas y la memoria local, produce una señal de salida y/o altera la memoria local.*
- *Las entradas a la RNA llegan del mundo exterior, mientras que sus salidas son conexiones que abandonan la RNA.*

Es posible encontrar y generar numerosas variantes de la anterior, utilizando grafos o evitando sus conceptos. Por lo general, se presta una atención especial al tratamiento de la información por parte del objeto definido. Así, una de las definiciones más completas y más utilizadas en la actualidad es la propuesta por el grupo de investigación *Parallel Distributed Processing Research Group* (pionero en la investigación en RNA):

Definición 2.4.11. *Una RNA está compuesta por los siguientes elementos básicos:*

1. *un conjunto de procesadores elementos o neuronas artificiales;*
2. *un patrón de conectividad o arquitectura;*
3. *una dinámica de activaciones;*
4. *una regla o dinámica de aprendizaje;*
5. *el entorno donde opera.*

Sin embargo, no se trata de una definición matemática propiamente dicha, sino que es más bien una enumeración de las partes que constituyen el ente a definir. De hecho, prácticamente ninguna de las definiciones encontradas es lo suficientemente precisa como para ser considerada definición en el sentido matemático. Es fácil encontrar "definiciones" que incorporan características de las RNA, pero es mucho más complicado describir las propiedades que debe verificar toda RNA y que no debe verificar nada que no sea una RNA. Partimos de la dificultad de que ni siquiera hay consenso en la comunidad científica sobre qué tipo de ente es una RNA (¿es un programa, un circuito electrónico, un grafo, un procedimiento estadístico...?); por otro lado, tampoco están claramente fijados los límites entre

lo que efectivamente es una RNA y lo que no debe ser denominado con el mismo nombre.

2.4.2. Propuesta de definición de RNA

Una vez realizada una revisión de la principales definiciones de RNA (algunas de las más relevantes se han enunciado en el apartado anterior), hemos decidido redactar una nueva definición, algo más precisa y que se ajuste a los distintos modelos de RNA existentes desde su origen hasta la actualidad. Somos conscientes de la dificultad de la tarea, pero creemos pertinente proporcionar las claves de nuestro intento. Introduciremos dichas claves progresivamente. Para comenzar, reescribiremos ligeramente la última definición planteada (Definición 2.4.11):

Definición 2.4.12. *Una RNA está constituida por los siguientes elementos básicos:*

1. *un conjunto de datos;*
2. *una estructura topológica subyacente;*
3. *unas familia de funciones neuronales;*
4. *una regla o dinámica de aprendizaje.*

Seguidamente, para expresar esta misma definición utilizando un lenguaje más propiamente matemático, se desarrollan cada una de las partes de lo que será la definición propuesta. Nótese que no se trata de un ejercicio exclusivamente teórico, sino que conocer las distintas características que debería tener una RNA puede ayudar a resolver problemas, esto es, diseñar la RNA que permita una mejor aproximación (o, al menos, una aproximación suficientemente buena) de la solución

perseguida. También utilizaremos dicha división para proponer posteriormente mejoras o modificaciones en cada una de las partes constituyentes de las RNA.

Conjunto de datos

Obviamente, como ocurre con el resto de partes de las RNA, hay diferentes formas de definir el conjunto de datos. Supondremos que los datos pueden ser expresados numéricamente (si es preciso, con una codificación apropiada). En cualquier caso, los datos están intrínsecamamente relacionados con el conjunto de variables que se pueden utilizar en la resolución del problema. Nótese que la naturaleza de estas variables puede ser binaria o continua (el resto de tipologías contemplables pueden trasladarse a uno de estos dos tipos, tal vez incrementando el número de variables binarias o sumergiendo las variables discretas en una continua), incluso pudieran combinarse ambos tipos en una misma RNA; esto no afecta sustancialmente a la definición que pretendemos dar del conjunto de datos.

Supongamos que se observan n características (o variables) de k individuos (u observaciones o casos). Para la observación i-ésima (con $1 \leq i \leq k$), se considera el vector $\overline{X}_i \in \mathbb{R}^n$, con n componentes que representan las n características observadas y potencialmente observables en los nuevos individuos; estas n características serán las utilizadas para inferir el comportamiento de los casos que puedan aparecer en el futuro.

En ocasiones, algunas otras características representan respuestas (o soluciones) que se pueden obtener al resolver el problema. Usualmente, estas características se interpretan como salidas (*outputs*) correspondientes a las entradas (*inputs*) incorporadas en cada vector $\overline{X}_i$ anteriormente descrito. Es decir, a cada $\overline{X}_i$ le correspondería otro vector $\overline{Y}_i \in \mathbb{R}^m$ con $1 \leq i \leq k$ (eventualmente vacío $\forall\, i$),

vector con m componentes ($0 \leq m$), donde estas m componentes representan las m características que se deseará conocer de los casos que se le puedan presentar a la RNA en el futuro.

Nótese que los valores del vector $\overline{Y}_j \in \mathbb{R}^m$ pueden ser desconocidos incluso para los individuos ya observados. En ese caso, $\overline{Y}_j$ lo supondríamos compuesto por incógnitas. Los números n (correspondiente a las variables de entrada u observables), m (correspondiente a las variables que se desea determinar) y $n+m$ constituyen diferentes formas de entender la dimensión del problema que se quiere abordar mediante una RNA.

Definición 2.4.13. *Sean k (casos), n (variables de entrada) y m (variables de salida) tres números enteros positivos. Llamamos* ***conjunto ordenado de datos*** *de una RNA al conjunto ordenado de k vectores de $\mathbb{R}^{n+m}$, donde las n primeras coordenadas de cada vector son necesariamente números conocidos mientras que las m restantes (para cada vector) son bien todos números conocidos o bien todas incógnitas.*

Para cada uno de los k vectores (i-ésimo caso, con $1 \leq i \leq k$), las primeras n de las $n+m$ coordenadas son las de $\overline{X}_i$ y las m siguientes son las de $\overline{Y}_i$. Así, el conjunto ordenado de datos de una RNA puede ser representado en forma de matriz (de dimensiones $k \times n+m$ o $n+m \times k$), o también en forma de conjunto ordenado de vectores (k vectores fila o vectores columna, según se convenga) del espacio vectorial $\mathbb{R}^{n+m}$.

Estructura topológica subyacente

Usualmente se define la estructura topológica de una RNA como la arquitectura que es necesaria para resolver el problema; dicha estructura viene determinada

por el número de neuronas y por las distintas conexiones existentes entre ellas. Por su naturaleza, la definimos con un grafo dirigido o digrafo $G = (V, E)$, de modo que los elementos de V son los vértices o nodos mientras que los elementos de E, que son pares ordenados de elementos de V y se denominan conexiones, arcos o aristas dirigidas. El número de arcos que llegan a un nodo se denomina *valencia de entrada* mientras que el número de arcos que salen de un nodo se denomina *valencia de salida*.

La estructura topológica depende, evidentemente, del conjunto de variables estudiadas. Así, cada variable de entrada corresponderá a un nodo con valencia de entrada nula, de los que hay n, según la notación utilizada en la Definición 2.4.13; análogamente, cada variable de salida corresponderá a un nodo con valencia de salida nula, de los que habrá m. Según esto, la estructura topológica también está relacionada con los datos, ya que ellos determinarán el conjunto de variables del problema y, consecuentemente, las neuronas imprescindibles en la RNA.

Además de los nodos correspondientes a las variables de los datos (de entrada o de salida), la RNA contará con un número determinado de nodos adicionales, que habitualmente se identifican con las neuronas. A cada conexión entre neuronas se le suele asociar un valor que denominaremos *peso*. La estructura de la red puede venir, por tanto, determinada por un vector o (lo que es más práctico) representada por una matriz de pesos, cuyas dimensiones dependerán del número de neuronas de la RNA. Si denotamos por W a la matriz de los pesos de la RNA, cada elemento de W, $w_{i,j}$, representa la influencia que tiene la neurona i sobre la neurona j. En general, los elementos $w_{i,j}$, puede ser positivos, negativos o nulos y su valor puede ser modificado con la fase de entrenamiento (descrita en los apartados posteriores de "Familia de funciones neuronales" y "Regla o dinámica de aprendizaje"). En el caso en que $w_{i,j} \equiv 0$ (es decir, $w_{i,j}$ no puede tomar un

valor distinto de cero), deberá prescindirse de la conexión entre la neurona i y la neurona j.

Aunque, salvo por la última consideración del párrafo anterior, la matriz W no forma parte de la estructura topológica subyacente, es interesante tenerla presente, pues será necesaria para la fase de entrenamiento (descrita en los dos apartados siguientes).

Definición 2.4.14. *Sea X un conjunto ordenado de datos de una RNA, según la notación utilizada en la Definición 2.4.13. Llamamos* ***estructura topológica subyacente*** *(de una RNA) asociada a X al digrafo (o grafo dirigido) etiquetado $G = (V, E)$, donde:*

- *V es el conjunto de neuronas (nodos del digrafo etiquetado).*
- *El arco de la i-ésima a la j-ésima neuronas está en E si y solo si existe la correpondiente conexión entre dichas neuronas (y, por tanto, existirá un peso $w_{i,j}$ no idénticamente nulo en la matriz de pesos W).*
- *Al menos n elementos de V tienen valencia de entrada 0; dichos elementos constituyen la llamada* capa de entrada.
- *Exactamente m elementos de V tienen valencia de salida 0; dichos elementos constituyen la llamada* capa de salida.

Aunque no se han establecido importantes restricciones topológicas hasta ahora (para facilitar el entrenamiento de la RNA, para posibilitar una definición lógica de familia de funciones neuronales y, como se verá enseguida, para establecer las capas o niveles de la RNA) sí es conveniente exigir que no existan caminos cerrados en el digrafo G. Es decir, por lo general, se evitará la presencia de una sucesión de arcos que empiecen y acaben en un mismo nodo.

Habitualmente, los nodos del digrafo G que no corresponden ni a la capa de entrada ni a la de salida se suelen clasificar en una o varias *capas ocultas*. Un modo sencillo de obtener dicha clasificación cuando no hay caminos cerrados en el digrafo es mediante un procedimiento recursivo sobre G: primero se suprimen los vértices de la capa de entrada (y los arcos que partían de dichos vértices); después se determinan los nuevos nodos con valencia de entrada nula y al conjunto constituido por estos nodos se les llama *primera capa oculta*; seguidamente se suprimen los nodos de la primera capa oculta (y los correspondientes arcos); y así sucesivamente se van determinando las restantes capas ocultas. Si el digrafo presenta algún camino cerrado, la determinación de las capas ocultas no es un ejercicio trivial y puede depender de los convenios que se adopten.

Como se deduce de la Definición 2.4.14, no es necesario considerar capas de salida con más de m elementos; sin embargo, sí es conveniente tener en cuenta que la capa de entrada puede tener más de n elementos. De hecho, independientemente de n y m, es frecuente añadir un nodo adicional por cada capa oculta del digrafo. Esos nodos adicionales se suelen llamar *unity bias* y para nosotros serán *entradas constantes independientes*.

Familia de funciones neuronales

Partiremos de una versión simplificada de la definición de función propuesta por Leibnitz:

Definición 2.4.15. *Una función es una regla que asigna a cada número (de entrada) exactamente un número (de salida). Al conjunto de números de entrada a los cuales se aplica la regla se le llama el dominio de la función, mientras que el conjunto de números de salida que se obtienen al aplicar la función es llamado*

el rango, co-dominio o recorrido.

Hay que tener en cuenta que los números de los que se habla en la definición anterior pueden pertenecer a diferentes tipos de dominio. Así, una función puede definirse sobre los números reales, sobre los enteros, sobre el $\{0, 1\}$, etc. De igual modo, el rango puede consistir en un conjunto finito o infinito de números, habitualmente reales en el ámbito que nos ocupa.

Ahora se trata de describir las funciones que formarán parte esencial de las RNA, algo que afectará a cada neurona y a cada conexión de la RNA. Y el número de dichos elementos dependerá del cardinal de V (esto es, $\#V$) y del cardinal de E (es decir, $\#E$). Específicamente, la dinámica de activación de cada neurona vendrá dada como la composición de dos funciones; a dicha composición la denominaremos *función neuronal* y a las funciones componentes las denominaremos *función de agregación* y *función de activación.*

Para poder expresar las definiciones siguientes, considérese una neurona que recibe información de s neuronas antecedentes (desde las que existe un arco hacia la propia neurona; dicha información procedente de cada una de las neuronas antecedentes también se llama estímulo o entrada) y que transmite (o emite) información hacia t neuronas subsecuentes o subsiguientes. Es decir, la valencia de entrada de la neurona sería s y la valencia de salida sería t.

Definición 2.4.16. *Una* **función de agregación** *incorpora la información de las s neuronas antecedentes. Cada estímulo se ve afectado (normalmente mediante un producto) por el peso de conexión entre la neurona antecedente y la actual (que se suele llamar peso sináptico); después se combina toda la información ponderada (frecuentemente de forma aditiva).*

Como es habitual, consideraremos que la ponderación se realiza mediante el

producto del valor que procede de una neurona antecedente por el peso correspondiente. Los tipos de funciones de agregación más comunes incorporarían las siguientes formas de combinación de información:

Suma: función lineal basada en la suma ponderada de las entradas por los pesos sinápticos asociados. Para cada neurona, se puede representar como el producto escalar del vector de entradas y el vector de los pesos. Constituye el tipo más habitual de función de agregación.

Multiplicación: función no lineal basada en la multiplicación entre las entradas (ya ponderadas) de la neurona.

Distancia euclídea: consiste en el cálculo de la norma euclídea del vector de entradas de la neurona (ya ponderadas sus coordenadas).

Distancia euclídea al cuadrado: consiste en elevar al cuadrado las entradas de la neurona (ya ponderadas sus coordenadas) y después sumarlas todas.

Cualquier otro tipo de media: consiste en elegir cualquier tipo de media o, incluso, de medida de tendencia central y aplicársela a las entradas de la neurona (ya ponderadas sus coordenadas).

Nótese que en todas las funciones anteriores, puede aparecer, como una componente más, la información procedente de la correspondiente entrada constante independiente. Muchos autores consideran este valor de forma separada y suele denotarse, para la capa j-ésima, por $b_{0,j} \times w_{0,j}$.

Definición 2.4.17. *Una* **función de activación o transferencia** *es una función monótona creciente y que se aplica sobre el resultado de aplicar una función de agregación.*

La función de activación es normalmente una función real de variable real $g : \mathbb{R} \to \mathbb{R}$, pero téngase en cuenta que las variables de la RNA pueden ser incluso binarias, por lo que la función de activación no iría de $\mathbb{R}$ a $\mathbb{R}$; en cualquier caso, lo habitual es definir la función en $\mathbb{R}$ y dejar abierta la posibilidad de restringirla al dominio discreto si fuera necesario.

A continuación se listan algunas de las funciones de activación más frecuentes en la práctica:

Función lineal o identidad: dada por $g(x) = x$, $\forall\, x \in \mathbb{R}$.

Función escalón o signo o limitador fuerte: dada por

$$g(x) = \begin{cases} 1 & \text{si } x \geq 0 \\ -1 & \text{si } x < 0 \end{cases} \quad \forall\, x \in \mathbb{R}$$

o bien

$$g(x) = \begin{cases} 1 & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases} \quad \forall\, x \in \mathbb{R}.$$

Función lineal a tramos o lineal a trozos: dada por

$$g(x) = \begin{cases} -1 & \text{si } x < 0 \\ x & \text{si } 0 \leq x \leq 1 \\ 1 & \text{si } x > 1 \end{cases} \quad \forall\, x \in \mathbb{R}$$

o bien

$$g(x) = \begin{cases} 0 & \text{si } x < 0 \\ x & \text{si } 0 \leq x \leq 1 \\ 1 & \text{si } x > 1 \end{cases} \quad \forall\, x \in \mathbb{R}.$$

Función sigmoidea: dada por $g(x) = \frac{1}{1+e^x}$, $\forall\, x \in \mathbb{R}$.

Función sigmoidal hiperbólica o tangencial: dada por $g(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$, $\forall\, x \in \mathbb{R}$.

Función tribas: dada por

$$g(x) = \begin{cases} 1 - |x| & \text{si } -1 \leq x \leq 1 \\ 0 & \text{en otro caso} \end{cases} \quad \forall\, x \in \mathbb{R}$$

Función gausiana o radbas: dada por $g(x) = \frac{1}{e^{x^2}}$, $\forall\, x \in \mathbb{R}$.

Función sinusoidal: dada por $g(x) = \text{sen}(x)$, $\forall\, x \in \mathbb{R}$.

Lo más frecuente es utilizar el mismo tipo de función de agregación y el mismo tipo de función de activación en todas las neuronas de una misma RNA o, al menos, en todas las neuronas de una misma capa de la RNA; sin embargo, no hay ninguna limitación teórica para permitir diferentes funciones componentes en cada neurona, más allá de la complejidad en el manejo de las expresiones resultantes.

Con las consideraciones presentadas hasta el momento, es razonable definir:

Definición 2.4.18. *Considérese una neurona cuya valencia de entrada es $s > 0$. Una* **función neuronal** *es una composición $g \circ f$, donde $f : \mathbb{R}^s \to \mathbb{R}$ es una función de agregación y $g : \mathbb{R} \to \mathbb{R}$ es una función de activación.*

A partir de la matriz de pesos W y de las funciones neuronales es posible definir las funciones clave para el posterior entrenamiento de las RNA:

Definición 2.4.19. *Sea G la estructura topológica subyacente de una RNA, con $\#V$ nodos y $\#E$ arcos, según la notación utilizada en la Definición 2.4.14. Llamamos* ***familia de funciones neuronales*** *asociada a la estructura topológica G de una RNA a función vectorial $F : \mathbb{R}^{n+\#E} \to \mathbb{R}^m$ tal que:*

1. *$F(x_1, x_2, \ldots, x_n, w_1, w_2, \ldots, w_{\#E}) = (y_1, y_2, \ldots, y_m)$; las primeras n variables (los x_i) corresponden a las entradas de la RNA; el resto de variables son en realidad parámetros que configuran la RNA y que se modifican durante el entrenamiento.*

2. *Una vez fijados los $\#E$ valores de los parámetros, cada función componente de F se puede obtener como la composición sucesiva de funciones neuronales.*

3. *En las composiciones a las que se refiere el apartado anterior de esta definición, una función neuronal se aplica sobre el resultado de otra si y solo si hay una conexión de la primera neurona hacia la segunda, según la configuración de G.*

Lógicamente, los valores w_i a los que se refiere la Definición 2.4.19 corresponden a los elementos de la matriz de pesos W correspondiente a G.

Nótese que la interpretación de la Definición 2.4.19 se complica sustancialmente si se permiten caminos cerrados dentro de G. En ese caso, será necesario establecer el número de composiciones máximo permitido a la hora de obtener la familia de funciones neuronales.

Regla o dinámica de aprendizaje

Como se ha introducido en la Definición 2.4.19, se precisa de un procedimiento para ir modificando progresivamente los pesos de W; es decir, se necesita una regla de aprendizaje que permita actualizar los pesos correspondientes a las conexiones entre las neuronas. Esta regla de aprendizaje viene habitualmente determinada por el tipo de datos que se tiene, por la arquitectura establecida para la RNA y por las funciones neuronales elegidas.

No siempre existe una "mejor" regla de aprendizaje, pero sí suele ser habitual que algunas reglas no sean suficientemente válidas, para el tipo de problema que se desea resolver, para la precisión que se desea alcanzar en la solución o para las

características que presentan el conjunto de datos X, la estructura topológica G o la función neuronal F.

Definición 2.4.20. *Sea F la familia de funciones neuronales de una RNA, según la notación utilizada en la Definición 2.4.19. Llamamos* ***regla de aprendizaje*** *asociada a la función F de una RNA a un algoritmo Λ que, a partir de un conjunto de datos apropiados (que se pueden expresar en una matriz real de dimensiones $k \times n$), permite obtener una solución (que se puede expresar en otra matriz real W de dimensiones máximas $\#V \times \#V$ o en un vector con $\#E$ componentes). Tal solución corresponde a los valores adecuados de los parámetros $\#E$ (w_i) de F.*

La solución a la que se alude en la Definición 2.4.20 es una que, con la precisión que se desee o que se pueda conseguir, permite que F resuelva el problema para el que se diseñó la RNA.

Habitualmente, las reglas de aprendizaje suelen dividir el conjunto de datos en dos partes: la primera (*conjunto de entrenamiento*) se utiliza para determinar los pesos mientras que la segunda (*conjunto de validación*) se usa para comprobar la bondad de la solución encontrada en cada paso o al final del proceso.

Retomamos ahora la Definición 2.4.12. Teniendo en cuenta las anteriores Definiciones 2.4.13, 2.4.14, 2.4.19 y 2.4.20, y siendo k, n y m tres números enteros positivos que describen la dimensión del problema a resolver, se define el concepto de RNA:

Definición 2.4.21 (Fedriani y Romano, 2014)**.** *Una RNA es una cuaterna (X, G, F, Λ), de modo que:*

1. *X es un conjunto ordenado de datos, según la Definición 2.4.13.*
2. *G es una estructura topológica subyacente asociada al conjunto ordenado de*

datos X, según la Definición 2.4.14.

3. *F es una familia de funciones neuronales asociada a la estructura topológica G, según la Definición 2.4.19.*

4. *Λ es una regla de aprendizaje asociada a la familia de funciones neuronales F, según la Definición 2.4.20.*

Llama la atención que la definición propuesta para RNA (Definición 2.4.21) comprende cuatro partes que participan de una estructura pluridisciplinar. Así, por una parte, se hace uso de Estadística o Informática; por otro lado, aparece la Topología o la Matemática Discreta; también se hace uso del Análisis Matemático, de los Métodos Numéricos y del Álgebra o la Computación. Finalmente, por su propio interés práctico, nos referimos a la Matemática Aplicada.

En cualquier caso, parece que los datos que se desean analizar y el problema que se pretende resolver son los que determinarán el resto de características de la RNA, algo que debe tenerse muy en cuenta cuando se diseñen redes apropiadas para una situación concreta o cuando se sugieran mejoras para los tipos de RNA ya ensayados.

2.5. Propiedades

A falta de definiciones precisas, las RNA se han descrito mediante su implementación específica o mediante sus propiedades. Así, por ejemplo, Hilera [68] propuso en el año 1995 un conjunto de propiedades que caracterizan a las RNA y que consideramos de interés para nuestros propósitos:

- **Aprendizaje adaptativo:** es la característica probablemente más atrac-

tiva de las RNA; a partir de casos conocidos, la RNA aprende a realizar las tareas mediante entrenamiento. Gracias a esta propiedad, no se necesita conocer o elaborar específicamente los modelos *a priori* sino que todo se obtiene a partir de su entrenamiento.

- **Autoorganización:** las RNA usan su capacidad de aprendizaje adaptativo para autoorganizar la información que recibe durante el aprendizaje y su posterior operación. Durante el proceso de aprendizaje se modifica cada elemento procesal y, de hecho, la autoorganización consiste en la modificación de la RNA completa para llevar a cabo un objetivo especifico.

- **Tolerancia a fallos:** las RNA son los primeros métodos computacionales con la capacidad inherente de tolerancia a fallos; comparados con los sistemas computacionales tradicionales (los cuales pierden su funcionalidad en cuanto sufren un pequeño error de memoria), en las RNA si se produce un fallo en un pequeño número de neuronas, aunque el comportamiento del sistema se vea influido, no sufre una caída repentina (esto es la tolerancia a fallos del propio sistema de computación). Incluso, las RNA pueden seguir realizando su función aunque se destruya parte de la red, porque tienen su información distribuida en las conexiones entre neuronas, existiendo cierto grado de redundancia en este tipo de almacenamiento. Puede decirse que las RNA almacenan información no localizada; por ello, la mayoría de las interconexiones entre los nodos de la red tendrá unos valores en función de los estímulos recibidos y se generará un patrón de salida que representa la información almacenada. Por otro lado, las RNA pueden aprender a reconocer patrones con ruido, distorsionados o incompletos (en esto consiste la tolerancia a fallos respecto de los datos).

- **Operación en tiempo real:** responde a una de las mayores prioridades de la mayoría de las áreas de aplicación: la necesidad de realizar grandes procesos con datos de forma muy rápida. Gracias a la implementación paralela, es posible llegar a realizar estas operaciones casi inmediatamente.

- **Fácil inserción dentro de la tecnología existente:** una RNA individual puede ser entrenada para desarrollar una única y bien definida tarea, independientemente de que dicha tarea forme parte de un proceso más complejo. Las RNA se pueden utilizar para mejorar sistemas de forma incremental y cada paso puede ser evaluado antes de acometer un desarrollo más amplio.

2.6. Tipos

Las RNA se pueden clasificar dependiendo de las distintas características comentadas anteriormente. Atendiendo a la Definición 2.4.21 y de forma natural, surgen varias posibles clasificaciones, aunque no todas tengan un sentido completo y coherente desde el punto de vista lógico-formal (de acuerdo a lo que es realmente una clasificación y para lo que puede servir).

2.6.1. Según el número de entradas o salidas

Las entradas son el equivalente a las variables independientes en los modelos clásicos, mientras que las salidas semejan las dependientes (o funciones que se quieren estimar o aproximar). Así, se puede hablar de RNA univariantes (una sola entrada), bivariantes (dos entradas), trivariantes (tres entradas)..., multivariantes (más de una entrada); también se pueden considerar RNA monoobjetivo (una sola salida) o multiobjetivo (más de una salida). Sin embargo, no existe un consenso

generalizado en las tipologías anteriores, pues una sola variable cualitativa puede requerir, por ejemplo, de la presencia de más de una neurona de entrada o de salida, sin que ello signifique que haya más variables en el problema que se está resolviendo.

Estas clasificaciones también podría decirse que dependen de la estructura topológica subyacente, pero es porque ciertas partes del grafo dependen de los datos utilizados.

Según el tipo de variables de entrada o de salida

Como ya hemos dicho, las variables de entrada de RNA pueden ser de varios tipos. Dichos tipos permiten una clasificación de las propias RNA. Algo similar puede hacerse con las variables de salida. Así, suelen distinguirse al menos dos tipos de RNA:

Binarias: si cada variable toma solo dos posibles valores, usualmente en $\{0, 1\}$ o en $\{-1, 1\}$.

Continuas: las variables pueden tomar, al menos de forma teórica, infinitos valores entre dos cualesquiera.

Nótese que, en la práctica, cualquier variable cuantitativa o cualitativa se puede expresar mediante combinaciones de variables binarias. En lo que respecta a las variables continuas, surgen como oposición a las discretas (o, en este caso, a las binarias), pero esto no quiere decir que la tecnología permita la utilización efectiva de infinitos valores entre dos cualesquiera.

2.6.2. Según la estructura topológica subyacente

En lo que sigue, trataremos de representar gráficamente ejemplos de RNA. Normalmente utilizaremos unos polígonos estrellados de 7 puntas para indicar la presencia de neuronas, aunque en ocasiones (por simplicidad) sustituiremos neuronas o conexiones por puntos suspensivos o, incluso, identificaremos las neuronas de la capa de entrada o de la de salida con las propias entradas y salidas, reduciendo así el tamaño y la complejidad del dibujo.

Según el número de capas ocultas

La arquitectura de la RNA ha sido frecuentemente utilizada para clasificar. En particular, el número de capas ocultas se ha mostrado considerablemente interesante para los investigadores.

Simple o monocapa: solo tiene una capa, aparte de las entradas y salidas.

Multicapa: tiene varias capas (aparte de entrada y salida), pero siempre un número finito de ellas.

Recurrente: existe algún camino cerrado entre dos neuronas; no es habitual que haya conexión entre dos neuronas de una misma capa; si alguna capa envía (parte de) su señal a alguna capa anterior, no es posible establecer propiamente el número de capas.

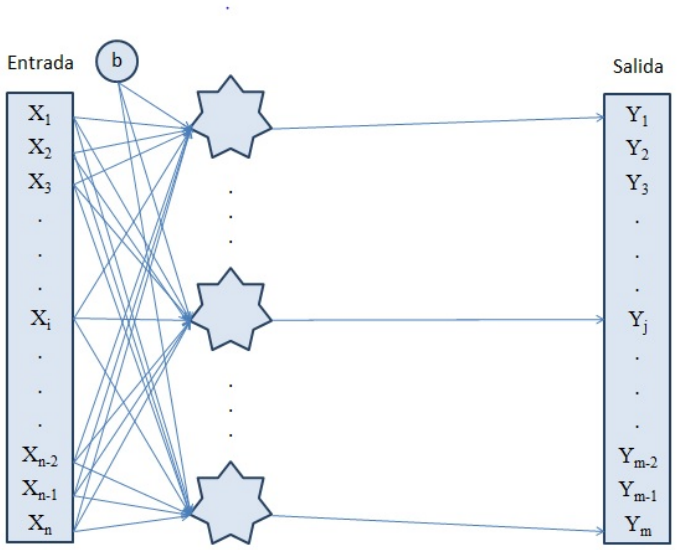


Figura 2.2: RNA monocapa

En rigor, en las RNA recurrentes el número de capas dependerá del número de iteraciones en el proceso de entrenamiento. Como el número de iteraciones no necesariamente es conocido *a priori*, esto complica ligeramente la interpretación de la Definición 2.4.19 (de familia de funciones de activación).

Conviene comentar que, en dichas RNA recurrentes, es habitual considerar que la capa que envía parte de su señal a una capa anterior es la capa de salida, pero según las definiciones presentadas anteriormente (al no poder tener valencia de salida positiva las neuronas de la capa de salida) se trataría de la capa justo anterior a la de salida (es decir, la última capa oculta). Análogamente, es habitual que la capa que recibe parte de la señal sea la capa de entrada, pero (al no poder tener valencia de entrada positiva las neuronas de la capa de entrada) se trataría de la capa siguiente a la de entrada (es decir, la primera capa oculta).

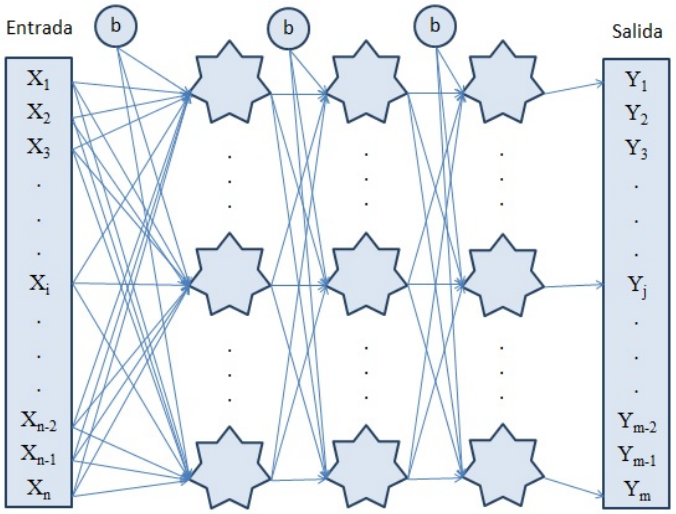


Figura 2.3: RNA multicapa

Según la conexión entre las capas

En los casos más típicos de RNA, las neuronas de una capa están exclusivamente conectadas con neuronas de la capa siguiente, pero en ocasiones es posible encontrar conexiones entre neuronas de una misma capa (creándose caminos cerrados dentro de una misma capa), entre neuronas de una capa y otra anterior (redes recurrentes) o entre una capa y otra posterior que no es la que la sucede inmediatamente. Esta casuística también es susceptible de ser adaptada para proporcionar una clasificación de las RNA (aunque no es algo frecuente).

Asimismo, es posible distinguir entre las RNA en las que todas las neuronas de una capa están conectadas con todas las neuronas de la capa siguiente (RNA completas) y las RNA en las que no se verifica dicha propiedad (RNA incompletas).

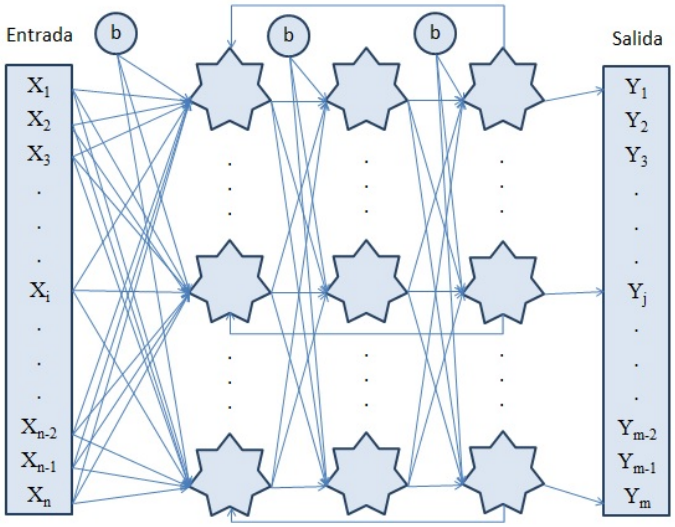


Figura 2.4: RNA recurrente

2.6.3. Según la familia de funciones neuronales

La existencia de diferentes tipos de funciones de agregación y de activación (como vimos antes de la Definición 2.4.19) hace posible la aparición de varias clasificaciones potenciales. Ha de tenerse en cuenta que la adopción de uno de estos tipos de funciones confiere a la RNA unas características esenciales (que justifican su relevancia) cuando la RNA se utiliza a efectos de clasificación, aproximación, estimación o simulación.

Función de activación

Algunas de la funciones de activación más utilizadas son:

- Limitador fuerte o función escalón

- Función lineal o identidad
- Función sigmoidal
- Función lineal a tramos o lineal saturada
- Función tangecial o función sigmoidal hiperbólica
- Función sinusoidal
- Función tribas
- Función gausiana

2.6.4. Según la regla de aprendizaje

Según el grado de conocimiento de las salidas

Como se comentó justo antes de la Definición 2.4.13, a veces se conocen las respuestas esperadas para unos datos de entrada, pero otras veces no es así y solo se conocen los propios valores de las variables de entrada en los casos observados. Esto provoca la aparición de diferentes sistemas de entrenamiento para la RNA, pues no es lo mismo evaluar la validez del ajuste mediante la comparación con resultados reales que sin ellos. Así, podemos distinguir al menos dos tipos de regla de aprendizaje (y, por tanto, de RNA):

Aprendizaje supervisado: la RNA conoce *a priori* las salidas correspondientes a las entradas presentadas o bien necesita que un experto valide los resultados que la propia RNA va proporcionando en el proceso de entrenamiento cada vez que se le presenta un conjunto de datos de entrada. En este tipo de aprendizaje, la RNA trata de minimizar el error entre la salida que calcula

y la salida deseada (conocida o valorada por el experto), de modo que la salida calculada termine siendo la deseada o tan próxima a ella como sea posible.

No supervisado o autoorganizado: la RNA es capaz de entrenar sin necesidad de conocer las salidas ni que un experto compruebe si las salidas parciales son satisfactorias. En ocasiones, no es fácil distinguir estos sistemas autoorganizados. Frecuentemente, la red conoce un conjunto de patrones, pero desconoce la respuesta deseada, con lo que debe extraer rasgos de los datos o agrupar patrones similares.

Según el entrenamiento

A lo largo de la historia de las RNA, los expertos han diseñado múltiples algoritmos con el fin de conseguir el entrenamiento más eficiente. Estos algoritmos pueden servir también para clasificar las propias RNA en las que operan:

Corrección de errores o minimizar el error: es el tipo de algoritmo más habitual, pues la RNA modifica sus pesos de una forma determinística cada vez que se comprueba que la salida proporcionada no es correcta. Dentro de este tipo se encuentran la reducción del gradiente, la retropopagación, etc. En cualquier caso, la modificación de pesos está siempre orientada a que el error cometido al final sea mínimo.

Estocástico o Regla de Boltzmann: se utilizan procedimientos estadísticos para modificar los pesos de forma pseudo-aleatoria. Se suelen aplicar en RNA estocásticas, donde se contemplan parámetros aleatorios.

Competitivo o cooperativo: son sistemas autoorganizados en los que no todas

las neuronas de salida se activan, sino solo las que presenten un mayor potencial sináptico. En estos casos, se podría decir que solo aprenden las neuronas que se acercan más a la salida deseada.

Hebb: se basa en que cuando el disparo de una célula activa otra, entonces el peso de la conexión entre ambas tiende a reforzarse (Ley de Hebb) [66].

Incluso sería posible clasificar las RNA según el tipo de problema que ayudan a resolver. En este caso, nos limitamos a dirigir al lector al apartado 2.8.

2.7. Ejemplos de RNA

A continuación se recogen algunos de los modelos de RNA más utilizados en la actualidad, observando algunas de sus características más relevantes.

2.7.1. Asociador lineal

Nombre: Asociador lineal

Datos: son conocidos los vectores X e Y, vectores de entrada y de salida, respectivamente; ambos vectores son definidos como de variables continuas.

Estructura topológica subyacente: es muy simple; solamente tiene la capa de entrada y la de salida, no existiendo capas ocultas (representada en la Figura 2.5).

Función neuronal: está compuesta por la función de agregación suma y la función de activación identidad, por lo que la función neuronal una función lineal.

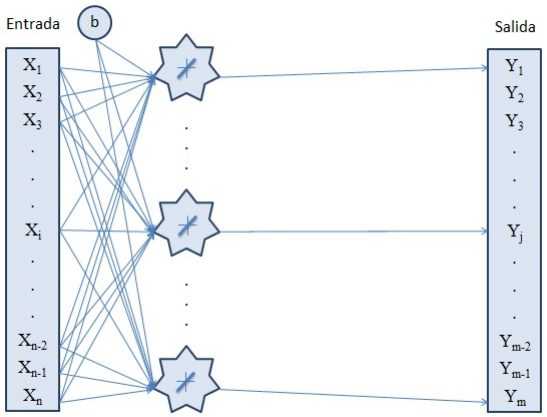


Figura 2.5: Asociador lineal

Dinámica de aprendizaje: el grado de conocimiento de los datos de salida es supervisado y el entrenamiento de la RNA está basado en la regla de aprendizaje de Hebb.

Aplicaciones: las principales aplicaciones consisten en asociar patrones.

2.7.2. Perceptrón simple

Nombre: Perceptrón simple; Rosenblatt [123], su creador en 1959, lo denominó *perceptron*; también se conoce como *perceptron network*.

Datos: los valores de entrada y salida son conocidos; los valores de entrada son continuos o binarios, mientras que los valores de salidas son siempre binarios.

Estructura topológica subyacente: este tipo de RNA es simple o monocapa, por lo que solamente posee la capa de entrada y la capa de salida y no existen capas ocultas (ver la Figura 2.6); existen conexiones entre las dos

capas de forma directa y de forma ascendente.

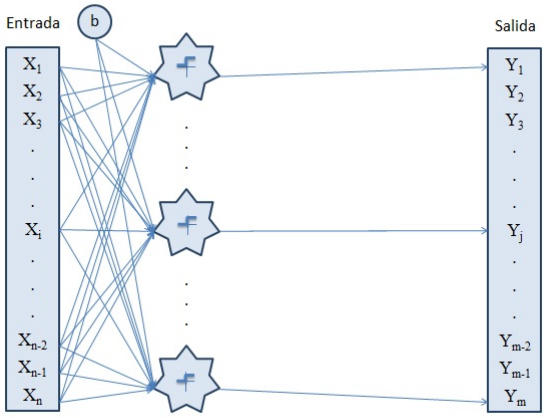


Figura 2.6: Perceptrón simple

Función neuronal: la función neuronal del percetrón, está compuesta por la función de agregación suma y la función de activación limitador fuerte o escalón, definida como: $g(x) = \begin{cases} 1 & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases} \quad \forall x \in \mathbb{R}$.

Dinámica de aprendizaje: el grado de conocimiento de los valores de salida es supervisado y el principal entrenamiento utilizado es la corrección de error o el de minimizar el error.

Aplicaciones: su principal aplicación es realizar tareas de clasificación, pero con la limitación de que las clases tienen que ser linealmente separables (para lograr un resultado satisfactorio).

2.7.3. Adaline

Nombre: *Adaptative Linear Element*, nombre propuesto por sus desarroladores, Widrow y Hoff, en el año 1960; también hay autores que la denotan por *Adalina*.

Datos: el vector de entrada es conocido y definido con valores continuos; el vector de salida es conocido, pero en origen estaba definido en los valores $\{-1, 1\}$, luego era de tipo binario; con posterioridad, se incorporaron redes Adaline con valores de salidas continuos (este cambio se produjo por la utilización de un nuevo algoritmo de entrenamiento).

Estructura topológica subyacente: es muy simple, solamente tiene la capa de entrada y la de salida, no existiendo capas ocultas (ver la Figura 2.7).

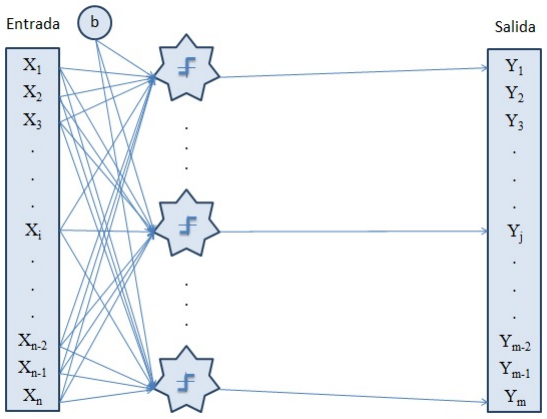


Figura 2.7: RNA Adaline

Función neuronal: está compuesta por la función de agragación suma y la función de activación limitador fuerte o escalón definida como: $g(x) =$

$\begin{cases} 1 & \text{si } x \geq 0 \\ -1 & \text{si } x < 0 \end{cases}$ $\forall x \in \mathbb{R}$, lo que algunos autores llaman conmutador bipolar.

Dinámica de aprendizaje: el grado de conocimiento de las salidas es supervisado y el principal entrenamiento utilizado originariamente era la corrección de error o regla del mínimo error cuadrado medio (LMS); posteriormente se empezó a utilizar la regla de Widrow-Hoff, convirtiendo la salida en valores continuos, para lo que se modifico la función de activación por una lineal.

Aplicaciones: las principales aplicaciones consisten en la eliminación del ruido en señales portadoras de información; también soluciona adecuadamente problemas que se puedan separar correctamente con patrones linealmente independientes.

2.7.4. Madaline

Nombre: *Multiple Adaline* (MADALINE); fue desarrollada por los mismos investigadores que la RNA Adaline y consiste en una sucesión de redes Adaline.

Datos: el vector de entrada está definido con valores continuos; el vector de salida es conocido, pero originariamente estaba definido con los valores $\{-1, 1\}$, luego era de tipo binario; con posterioridad se utilizó otro algoritmo de aprendizaje, convirtiendo la salida en continua.

Estructura topológica subyacente: es una sucesión ordenada de redes Adaline (ver la Figura 2.8).

Función neuronal: es la misma que utiliza cada RNA Adaline.

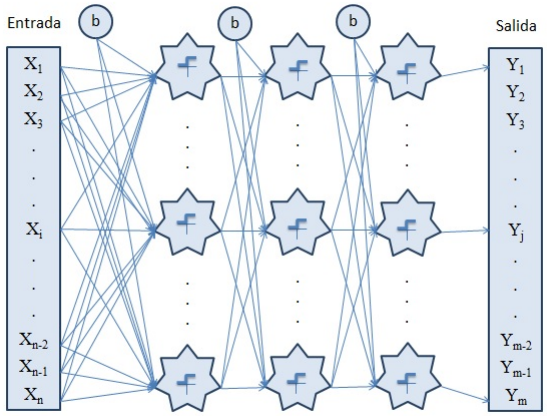


Figura 2.8: RNA Madaline

Dinámica de aprendizaje: el grado de conocimiento de las salidas es supervisado y el principal entrenamiento utilizado es como el de la RNA Adaline.

2.7.5. Perceptrón multicapa (PM)

Nombre: también se suele llamar *feed-forward* o *Multi-Layer Perceptron* (MLP).

Datos: los valores de entradas son continuos y la salida es conocida y definida con valores continuos.

Estructura topológica subyacente: la estructura mínima establecida es una capa de entrada, una capa oculta y una de salida (ver la Figura 2.3), pudiendo ser el número de capas ocultas cualquier número finito, aunque normalmente se presentan una o dos capas ocultas.

Función neuronal: está compuesta por la función de agregación suma como la principal utilizada y la función de activación más útil en este tipo de red es

la función sigmoide.

Dinámica de aprendizaje: el grado de conocimiento de las salidas es supervisado y el principal algoritmo de entrenamiento utilizado es la corrección de error o minimizar el error. El nombre utilizado para denotar al PM depende mucho del tipo de entrenamiento utilizado; así, también es conocido como RNA feed-forward. El algoritmo de aprendizaje más utilizado es el denominado retro-propagación o *back-propagation*, implementado por Rumelhart y Parker en 1985 ([127] y [109]); este tipo de aprendizaje realmente está basado en minimizar la función de error mediante el método de descenso de gradiente conjugado que años después fue ampliado por Battiti [9]. Algunos de los algoritmos de entrenamiento más eficientes en la actualidad se conocen como *Extreme Learning Machine* (ELM), aunque también hay quien los denota como MLP (por las propias siglas en inglés del PM). También destaca la RNA denominada *Quickpropagation*, que se caracteriza por su algoritmo *Quickprop*, que fue desarrollado por Fahlman en 1988 [34]. Otro de los tipos de aprendizaje más usuales es por la correlación en cascada, método desarrollado por Fahhman y Lebiere en 1990 [35].

Aplicaciones: la aplicación más destacada de las redes MLP se basa en la propiedad comentada por Kolmogorow en 1957 y, posteriormente, afirmada por autores como Funahashi, Hecht-Nielsen, Sarle y White, por la cual la RNA PM se puede utilizar como aproximador universal de funciones (para una información más detallada de la propiedad y su demostración, se puede consultar [25], [74] y [75]).

2.7.6. Función de base radial

Nombre: este tipo de red fue propuesto inicialmente por Powell en 1987, con el nombre de *estadísticos convencionales*; posteriormente recibió el nombre de *radial basis function* (RBF).

Datos: son conocidos los vectores de entrada y salida; ambos vectores son de valores continuos.

Estructura topológica subyacente: la estructura de la RBF es muy simple y se organiza en tres capas de neuronas: la capa de entrada, la capa oculta y la capa de salida (ver la Figura 2.9).

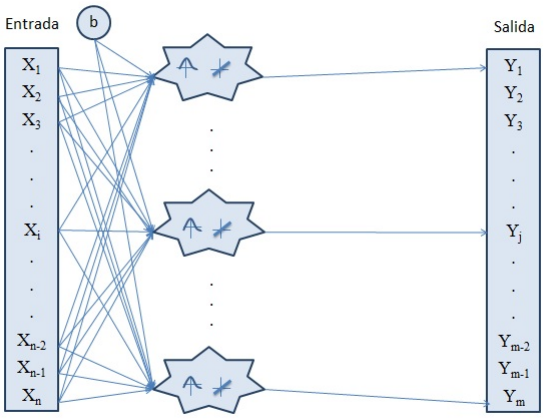


Figura 2.9: Función de base radial

Función neuronal: en su formato más usual, está compuesta por la función de agragación suma y la función de activación gausiana.

Dinámica de aprendizaje: es una red de tipo supervisado y el algoritmo de aprendizaje es autoorganizado (competitivo o cooperativo).

Aplicación: las aplicaciones más usuales son la realización de predicciones y clasificaciones; se utiliza frecuentemente como aproximador local de una función no lineal. Este tipo de RNA, junto con el PM, es de los más utilizados en la práctica.

2.7.7. Red de Hopfield

Nombre: denomina *Hopfield network* o *Discrete Hopfield*, al ser desarrollado por Hopfield en 1982.

Datos: son conocidos los vectores de entrada y salida; los valores de entrada tiene un carácter continuo y la salida un carácter binario: $\{-1, 1\}$ o $\{0, 1\}$.

Estructura topológica subyacente: está compuesta por tres capas: una de entrada, una de salida y una única capa oculta, donde todas las neuronas de la capa oculta esta interconectadas (ver la Figura 2.10).

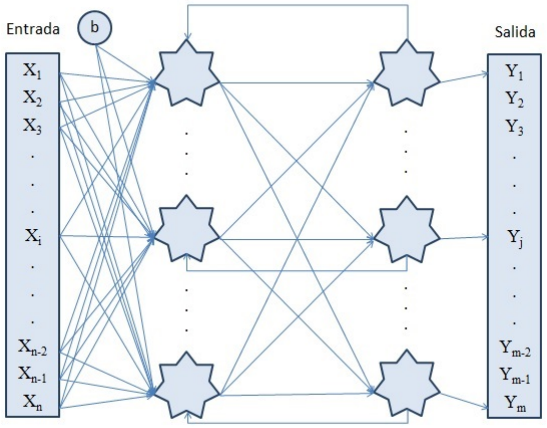


Figura 2.10: Red de Hopfield

Al existir conexiones entre neurones y por propia definición no existe interconexión entre las neuronas de una misma capa, se añadirá capa, todas ellas iguales, igual al número de iteraciones necesarias para llegar a la solución deseada.

Función neuronal: está compuesta por la función de agregación suma y la función de activación escalón.

Dinámica de aprendizaje: el aprendizaje es supervisado y se aplica el algoritmo de aprendizaje de la regla de Hebb.

2.7.8. Red continua de Hopfield

Nombre: se desarrolló como la versión continua de la RNA Hopfield, en el año 1984; también se denomina *Hopfield continuous network* o *Continuous Hopfield.*

Datos: son conocidos los vectores de entrada y de salida; tanto los valores de entrada como los de salida son de carácter continuo.

Estructura topológica subyacente: está compuesta por tres capas: una de entrada, una de salida y una única capa oculta, donde todas las neuronas de la capa oculta están interconectadas (ver la Figura 2.10).

Función neuronal: está compuesta por la función de agregación suma y la función de sigmoide o tangente hiperbólica.

Dinámica de aprendizaje: el aprendizaje es supervisado y se aplica el algoritmo de aprendizaje de la regla de Hebb.

2.7.9. Learning Vector Quantization Network

Nombre: también denotada como LVQ, se desarrolló desde 1980 por parte de Linde y Gray, aunque comenzó a utilizarse como herramienta para la compresión de datos en 1992.

Datos: son conocidos los valores de entrada y las clases en que se desea clasificar (lo que, según nuestras definiciones, forma parte de la dinámica de aprendizaje).

Estructura topológica subyacente: se caracteriza por una estructura multicapa, organizada por una capa de entrada, una o varias capas ocultas y la capa de salida.

Función neuronal: este tipo de red utiliza como función de agregación la distancia euclídea y como función de activación la identidad.

Dinámica de aprendizaje: aunque solo se conocen los valores de entrada, se caracteriza por tener un aprendizaje supervisado, ya que utiliza la clases para mover los denominados *vectores de Voronoi*; su aprendizaje es competitivo.

Aplicaciones: clasificador; el clasificador LVQ siempre divide el conjunto de datos de entrada en clases disjuntas.

2.7.10. Red probalística

Nombre: *Probalistic neural network*; también se nombra por sus siglas en inglés, PNN.

Datos: son conocidos los datos de entrada y de salida; los valores de entrada son binarios y también lo son los de salida.

Estructura topológica subyacente: está compuesta por cuatro capas: una de entrada, dos ocultas y una de salida.

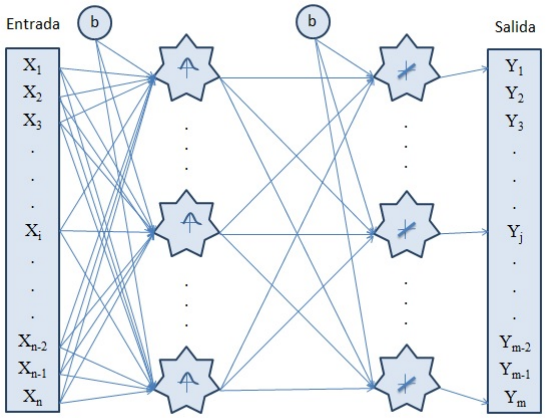


Figura 2.11: Red probabilística

Función neuronal: la función de agregación que se utiliza es la media, para normalizar el rango de los valores; la función de activación es la función gausiana en la primera capa oculta y la función de activación lineal en la segunda capa oculta.

Dinámica de aprendizaje: es supervisada, con un algoritmo de aprendizaje de minimizar el error.

Aplicación: se caracteriza por realizar una clasificación cercana a la clasificaciones de Bayes óptima; por eso, se utiliza para la clasificación de patrones utilizando métodos estadísticos.

2.7.11. Red de regresión general

Nombre: *General regression neural newtork*, denotada habitualmente por GRNN.

Datos: conocidos los datos de entrada y salida; los valores de entrada son binarios mientras que los datos de salida son continuos.

Estructura topológica subyacente: está compuesta por cuatro capas: una de entrada, dos ocultas y una de salida.

Función neuronal: está compuesta por la función de agregación suma y la función de activación no lineal.

Dinámica de aprendizaje: la dinámica utilizada es supervisada, con un algoritmo de aprendizaje de minimizar el error.

Aplicación: se caracteriza por realizar una clasificación cercana a la clasificaciones de Bayes óptima.

2.7.12. Brain State in a Box

Nombre: esta red fue desarrollada por Anderson y otros investigadores en 1972, como una aproximación a la red de tipo Hopfield.

Datos: son conocidos los vectores de entrada y de salida; tanto los valores de entrada como los de salida son de carácter continuo.

Estructura topológica subyacente: está compuesta por una capa de entrada, una o varias capas ocultas y una capa de salida, donde el número de neuronas de la capa de salida viene determinado por la dimensión de los patrones que se desea memorizar.

Función neuronal: la función de agregación utilizada es la función suma y la función de activación es la función lineal a trozos.

Dinámica de aprendizaje: el grado de conocimiento de la salida es supervisado y el principal entrenamiento utilizado es la corrección de error o minimizar el error.

Aplicación: su principal aplicación es en tareas de autoasociación.

2.7.13. Mapa autoorganizado

Nombre: se denotan también como *self organizing maps* (SOM) o bien *self-organizing fearture Map* (SOFM). Algunas redes de este tipo son conocidas con otros nombres, como el mapeo de diferentes dimensiones y las redes híbridas.

Datos: solo son conocidos los valores de entrada, que se caracterizan por ser binarios.

Estructura topológica subyacente: está compuesta por dos capas: una de entrada y otra de salida, como se puede observar en la Figura 2.12; la red es de tipo unidireccional.

Función neuronal: la función de agregación utilizada es la distancia euclídea o distancia euclídea al cuadrado.

Dinámica de aprendizaje: el tipo de aprendizaje es de tipo no supervisado; su algoritmo de aprendizaje es autoorganizado.

Aplicaciones: su principal aplicación es el reconocimiento de patrones y el procesamiento de imágenes, mediante la organización de mapas topológicos.

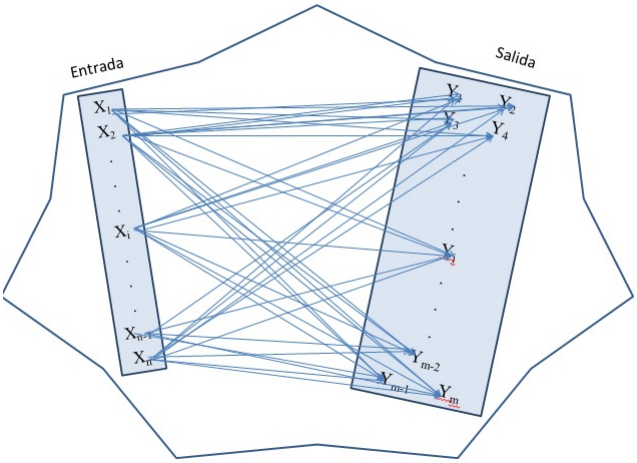


Figura 2.12: Mapa autoorganizado

2.7.14. Máquina de Boltzmann

Nombre: El nombre más usual es *The Boltzmann Machine*, propuesto por Hinto en 1984; ha sufrido ligeras modificaciones generando la *Boltzmann Feedforward*, la *Boltzmann Input-Output* o la *Boltzmann Completion Network*.

Datos: los datos de entrada son continuos y los datos de salida son binarios.

Estructura topológica subyacente: la red es de tipo multicapa y las capas están conectadas bidireccionalmente.

Función neuronal: es muy caracteristica; se trata de la denominada "función de Fermi" [101], que es una función de probabilidad.

Dinámica de aprendizaje: se denomina precisamente "entrenamiento de la máquina de Boltzmann", que es supervisado; se trata de una de las redes de tipo estocástico más desarrollada.

Aplicaciones: realizar tareas de auto y heteroasociación de patrones.

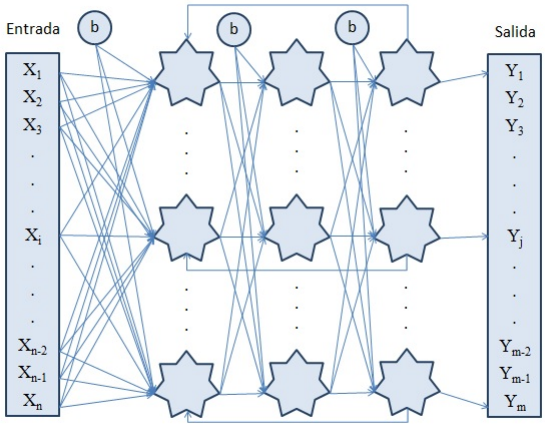


Figura 2.13: Máquina de Boltzmann

2.7.15. Adaptive Resonance Theory Network

Nombre: se suelen denotar por las siglas ART; dentro del conjunto de ART, existen distintas RNA.

Datos: solamente son conocidos los valores de entrada, que están expresados en forma binaria.

Estructura topológica subyacente: existe una capa de entrada y una de salida, como se puede apreciar en la Figura 2.14.

Función neuronal: la función de activación es la identidad, mientras que la función de agregación es una función de distancia.

Dinámica de aprendizaje: estos tipos de modelos tienen un aprendizaje no supervisado, son capaces de formar grupos o categorías a partir de secuencias de patrones de entrada y sin conocer su salida; dependiendo de si se utilizada una RNA de tipo 1 o de tipo 2, en la primera solamente devuelve la

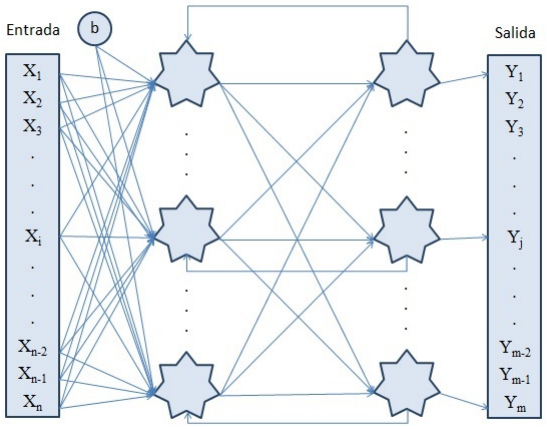


Figura 2.14: RNA adaptive resonance theory network

agrupación mientras que en el tipo 2 devuelve también el número de grupos que lo configuran; las redes de tipo 1 se denominan ART1 y las de tipo 2 ART2.

Aplicaciones: su principal aplicación es el reconocimiento de patrones y modelar el sistema neuronal.

2.7.16. Memoria asociativa bidireccional

Nombre: *Bidirectional associative memory* (BAM), desarrollada por Bart Kosko en 1987.

Datos: solo son conocidos los valores de entrada, en este caso valores binarios.

Estructura topológica subyacente: la estructura es multicapa, con capa de entrada conectada a la capa oculta y esta a la capa de salida; las conexiones son bidireccionales (ver la Figura 2.15).

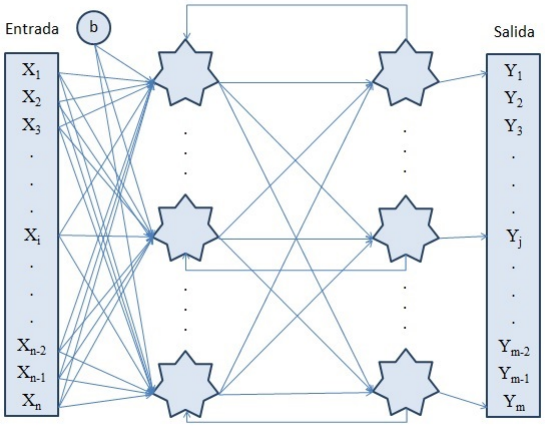


Figura 2.15: RNA de memoria asociativa bidireccional

Dinámica de aprendizaje: utiliza la mismo que las de la red de tipo Hopfield; el aprendizaje es supervisado y aplicando el algoritmo de aprendizaje de la regla de Hebb.

Aplicaciones: principalmente, de asocación; por ejemplo, asociar a un código binario de 10 bits, una firma digitalizada de 10.000 bits, o a una imagen de 140.000 bits una imagen comprimida de 7.000 bits.

2.7.17. Máquina de Cauchy

Nombre: la máquina de Cauchy (o *Cauchy Machine*) es una versión mejorada de la máquina de Boltzmann.

Datos: los datos de entrada son continuos y los datos de salida son binarios.

Estructura topológica subyacente: la red es de tipo multicapa, con capas conectadas bidireccionalmente (ver Figura 2.13).

Función neuronal: es muy característica; se trata de la denominada "función de Cauchy", que es una función de probabilidad; esta función converge con mayor rapidez que la utilizada en la maquina Boltzmann, siendo esta la mayor diferencia entre ambas RNA.

Dinámica de aprendizaje: se caracteriza por el entrenamiento de la máquina de Boltzmann, siendo supervisado; se trata de una de las redes de tipo estocástico más desarrollada.

Aplicaciones: realizar tareas de auto y heteroasociación de patrones.

2.7.18. Mapa de Kohonen

Nombre: mapas o redes de Kohonen; en 1982 Kohonen presentó este modelo sencillo para la formación autoorganizada de mapas de características de los datos de entrada.

Datos: solamente son conocidos los datos de entrada y, en general, son de tipo binario.

Estructura topológica subyacente: es muy simple, con capa de entrada y capa de salida, aunque en la capa de salida la distribución es distinta, como se puede apreciar en la Figura 2.12.

Función neuronal: la función neuronal más utilizada es la función a trozos, donde se determinan los valores en que se clasificarán los datos de entrada, mediante cálculos de distancias; esta regla de aprendizaje se denota como *mapa autoorganizativo*.

Dinámica de aprendizaje: es un proceso de aprendizaje no supervisado y trata de asociar características o patrones significativos en los datos de entrada.

Aplicaciones: en general, se utiliza para la visualización de datos.

2.7.19. Counter Propagation Network

Nombre: también denominada red *Counterpropagation*, que fue desarrollada por Rober Hecht-Nielsen en 1987.

Datos: los vectores de entrada y de salida son conocidos y con carácter binario o continuo.

Estructura topológica subyacente: la estructura más usual es la compuesta por tres capas: entrada, oculta y salida.

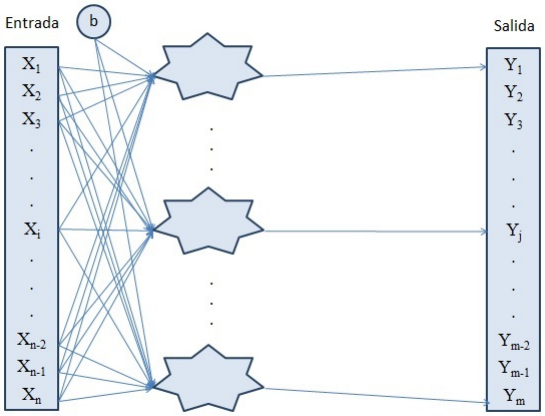


Figura 2.16: Counter Propagation Networks

Función neuronal: está compuesta por la función de agregación suma y la función de activación lineal.

Dinámica de aprendizaje: se caracteriza por utilizar dos tipos de aprendizaje: uno en la capa oculta, utilizando un algoritmo competitivo y no supervisado,

y otro en la siguiente capa, capa de salida, que no es competitivo y utiliza la regla de Widrow-Hoff, siendo supervisado.

Aplicaciones: algunas de las más usuales son la clasificación de patrones, la aproximación de funciones, el análisis estadístico y la compresión de datos.

2.7.20. Red neuronal competitiva no supervisada

Nombre: RNA competitivas no supervisadas

Datos: los valores conocidos son los que determinan el vector de entrada y están definidos de forma binaria.

Estructura topológica subyacente: estáa compuesta por dos capas: la capa de entrada y la capa de salida.

Función neuronal: la función de agregación más utilizada es el valor máximo y la función de activación suele ser la función limitador fuerte.

Dinámica de aprendizaje: sistema autoroganizado, luego la dinámica del aprendizaje es no supervisado y el algoritmo de aprendizaje es competitivo.

Aplicación: la principal es resolver problemas de clasificación.

2.8. Aplicaciones

Jonas Sjöberg [133] distingue entre cuatro tipos de aplicaciones para las RNA: clasificación, aproximación, estimación o simulación. Aunque algunos problemas pueden ser considerados en dos o más de estos tipos, estimamos que es un buen punto de partida para ordenar las aplicaciones más relevantes que se han sucedido

en los últimos años. Así, las RNA se pueden utilizar para el reconocimiento de patrones, para la compresión de información y la reducción de la dimensionalidad, para el agrupamiento o la clasificación, para la visualización, etc. Por sus características operativas, es una herramienta muy utilizada para cualquier problema relacionado con la minería de datos; recientemente, está siendo utilizada con éxito en el diagnóstico de enfermedades, para reconocer imágenes o, en general, para aproximar funciones en diferentes ámbitos.

A continuación se enumeran un par de las aplicaciones más llamativas que se han realizado en los últimos años en varios campos de conocimiento muy distintos entre sí. No es posible dar referencias de publicaciones científicas en todos los casos debido a que el interés de los diseñadores de las RNA a menudo está reñido con la publicación de las herramientas utilizadas y de los resultados obtenidos.

- Sobre sistemas educativos:
    1. Clasificar los niveles de calidad de un determinado sistema educativo utilizando una RNA [16].
    2. Generar sistemas de enseñanza-aprendizaje automático que se adapten al alumno.
- Para la imputación de datos faltantes:
    1. Reposición de los datos en la Seguridad Social de UK para rellenar los datos en blanco de las encuestas anuales; en una ocación se rellenó (posteriormente) el 70 % de los 45000 campos que se encontraban en blanco.
    2. Determinación de perfiles de usuarios o clientes con información incompleta (para ofrecer seguros adecuados o para realizar neuro-marketing).

- En Medicina:
    1. Detección de células cancerosas; desde el año 2009 se emplean RNA para la detección y para el tratamiento de dolencias de naturaleza cancerígenas [18].
    2. Detección de lesiones neurológicas y cardíacas.
- En Lingüística:
    1. Entrenamiento de sistemas automáticos de traducción.
    2. Determinación del ganador de un debate político, herramienta que posibilita el entrenamiento posterior de oradores [51] y [49].
- En Informática:
    1. Creación de vídeojuegos que simulan la existencia de un jugador humano (algunos ejemplos conocidos son Quake II o varios programas de ajedrez).
    2. Reconocimiento de voz o de escritura humanas.
- En el ámbito de la naturaleza:
    1. Patrones para el seguimiento de diversas especies migratorias de aves, anticipando su posición desde diversas variables relacionadas con la condiciones principales climatológicas de su medio.
    2. Prediciones meteorológicas, mediante el seguimiento de la ubicaciones de borrascas y periodos anteriores y anticiclónicos.
- En aviación:
    1. Determinación de la asignación de pistas de aterrizaje y despegue más apropiadas.

2. Reconocimiento de rostros de personas potencialmente peligrosas en aeropuertos.

A continuación se presentan algunas aplicaciones más, relacionadas especialmente con el ámbito económico, por guardar más relación con el ámbito en que se desarrolla la presente memoria.

2.8.1. Aplicaciones en el ámbito económico o empresarial

Los principales problemas que se han resuelto mediante RNA en el ámbito económico son de dos tipos, fundamentalmente: de clasificación o de aproximación de funciones, aunque dentro de estos últimos destaca el subgrupo de los problemas de simulación. A continuación comentaremos brevemente algunos ejemplos destacados de los problemas resueltos.

Problemas de clasificación

- Predicción del fracaso empresarial y del riesgo de crédito de deudores.
- Detección de fraude en el uso de tarjetas de crédito.
- Decisión de concesión o no de préstamos a solicitantes.
- Calificación o *rating* de obligaciones.
- Identificación de segmentos de mercado.
- Predicción de la opinión de los auditores.
- Elección del método de gestión de almacenes.

En la mayoría de los problemas de clasificación anteriores, las soluciones obtenidas se compararon con las aportadas por técnicas tradiciones (nos referimos, en particular, a regresión lineal, análisis discriminante, análisis logit y análisis probit, entre otros). Como se sabe, por su propia definición, el rendimiento de la RNA depende de la arquitectura neuronal elegida (topológica y funcional), de los parámetros y el aprendizaje seleccionados... y no siempre es fácil encontrar la RNA más apropiada para un problema. En cualquier caso, a pesar del inconveniente anterior, se observa que las redes de tipo perceptrón multicapa (MLP) superan ampliamente a los modelos tradicionales, aunque con menor diferencia cuando se comparan con los métodos analíticos logit y probit.

Problemas de aproximación de funciones

- Análisis financiero y bursátil.
- Análisis técnico relativo a la predicción de la cotización de acciones.
- Análisis de los vectores y series temporales para un determinado proceso, buscando los ritmos de producción más idóneos [131, 137, 7].
- Evaluación de las curvas de consumo de un determinado producto, lo cual permite predecir los óptimos de producción [73].
- Determinación de las oscilaciones del precio de los bienes [17].
- Búsqueda de la eficiencia en la cadena productiva [110].
- Predicción de los movimientos de la tasa de actividad, anticipando posibles cambios en el desempleo [107].
- Valoración de las variables de tipo macroeconómico y cómo condicionan el rendimiento empresarial [86, 135, 134].

- Evolución de los diversos tipos de cambio [81, 143, 54].
- Predicción de las futuras fluctuaciones [115].
- Valoración y previsión del consumo de energía eléctrica de carga para disponer de las instalaciones más idóneas para la demanda que se presente [24, 83].
- Mediciones eléctricas [130].
- Seguimiento y previsión de la evolución de economías dinámicas [19].
- Análisis bursátiles, de cara a encontrar las variables que condicionan las oscilaciones de los mercados [76, 120, 64, 67, 52].
- Valoración de las fluctuaciones del empleo, en función de la estación del año y también de las características intrínsecas de cada país [53].
- Análisis del control del gasto público en función de las incapacidades laborales [122].
- Interpretación de las variables en materia de economía regional, para calibrar los efectos de la Unión Europea sobre la industria europea [102].
- Análisis de riesgos en entidades financieras [103].
- Aumento y disminución de la deuda internacional, considerando factores internos y externos, y la relación comercial entre los diversos estados [21].
- Predicción de los diversos niveles de inflación y cómo estos condicionarán el consumo, sin necesidad de modelizar la curva de los precios [100].
- Evolución de las tasas de crecimiento de una empresa o sociedad, teniendo en cuenta los diversos elementos del tejido productivo [139].

- Optimización en tablas input-output, superando las técnicas econométricas tradicionales [108].
- En el mundo del Marketing, para la elección de una marca, realizando una predicción (simulación) sobre la acogida que presentará en los potenciales usuarios o compradores [144].
- Creación de indicadores para cuantificar la pobreza de un país, posibilitando la comparación entre diferentes estados, así como estudio de otros conflictos macroeconómicos (sugerido en [89, 41]).
- Estimación de la prima de Riesgo y modelo predictivo para los ingresos a partir de la Encuesta Nacional de Ingresos y Gastos de los Hogares [140].

Capítulo 3

Propuestas de ajustes de la técnica

3.1. Tratamiento de bases de datos incompletas

3.1.1. Preliminares

La Informática surgió, como su propio nombre indica, para tratar con la información; en esto, guarda una estrecha relación con la Estadística. Obviamente, la información puede ser de muchos tipos distintos y no siempre es apta para aplicar la metodología deseada por el investigador, bien por los requerimientos técnicos (informáticos) o bien por el incumplimiento de las hipótesis teóricas (estadísticas). De hecho, uno de los grandes problemas que muchos investigadores de distintas disciplinas se plantean desde hace años es la falta de información adecuada y fiable, así como la pérdida de datos correspondientes a las variables que se desean estudiar.

Al inicio de los años 70 del siglo XX distintos autores, como Afifi y Elashoff [2] o Hartley y Hocking [65], publicaron los primeros tratados sobre los datos incompletos. Desde entonces numerosos investigadores han desarrollado diversas técnicas para el tratamiento de los valores perdidos en distintos tipos de bases de datos. A modo de ejemplo reciente y cercano a nuestro grupo de investigación, la tesis de la Profesora Sánchez [129] trata sobre el desarrollo de técnicas alternativas a las habituales para paliar el problema de los datos perdidos en series temporales.

En el siguiente apartado se realiza una revisión de algunos tratamientos posibles ante el problema de los datos perdidos, para luego presentar una nueva técnica para incorporar los datos perdidos en trabajos de investigación, apoyándose en las RNA.

3.1.2. Introducción al problema de clasificación con datos perdidos

A fin de presentar la nueva metodología y de compararla en lo posible con las técnicas ya existentes, se utiliza un lenguaje algo más cercano al formal que al natural. Así, se considera el siguiente problema:

Sea una población finita formada por N individuos, donde se denotará por X_i al individuo i-ésimo ($i = 1, 2, \ldots, N$). De estos individuos, idealmente, se conoce el valor que alcanzan para M variables distintas. Cada una de estas variables viene expresada por Y_j: la j-ésima observación del conjunto de individuos. Así, de forma resumida, el conjunto $\{X_1, X_2, ..., X_N\}$ es el formado por los N individuos e $\{Y_1, Y_2, ..., Y_M\}$ representa el conjunto de las variables estudiadas. Supongamos que se desea realizar una clasificación (de cualquier tipo) o bien una ordenación de los N individuos teniendo en cuenta las M variables conocidas.

Para simplificar la notación posterior, se denota por $a_{ij} = Y_j(X_i)$ a la observación de la variable Y_j sobre el individuo X_i. Coherentemente, la matriz de datos reales multivariantes formada por los a_{ij} se denotará como $A \in \mathcal{M}_{N\times M}(\mathbb{R})$:

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1j} & \cdots & a_{1M} \\ \vdots & \ddots & \vdots & & \vdots \\ a_{i1} & \cdots & a_{ij} & \cdots & a_{iM} \\ \vdots & & \vdots & \ddots & \vdots \\ a_{N1} & \cdots & a_{Nj} & \cdots & a_{NM} \end{pmatrix},$$

donde la fila i-ésima de A representa las M variables observadas del individuo X_i y la columna j-ésima de A representa el resultado de observar la variable j-ésima a los N individuos. Luego el vector $(a_{i1}, \ldots, a_{ij}, \ldots, a_{iM})$ representa las observaciones de las M variables sobre el individuo X_i.

En primer lugar, hemos de reconocer que si se desea estudiar el comportamiento de los N individuos atendiendo a las M variables, es recomendable realizar un estudio preliminar sobre la información y características de los datos con los que se cuenta; es decir, dicho análisis preliminar se debe ejecutar antes de efectuar la clasificación u ordenación que se persigue finalmente.

En el caso de existir todos los valores a_{ij}, de comprobar que son correctos y que están bien definidos, no debería existir ningún problema a la hora de analizar la relación entre las variables y los individuos estudiados, así como de llegar a clasificarlos correctamente (u ordenarlos, si es eso lo que se pretende lograr) mediante cualquiera de las técnicas usuales para ello.

Pero aquí nos planteamos qué ocurre si existen datos perdidos: valores a_{ij} (o elementos de la matriz A) que no son conocidos, que realmente no existen o que,

existiendo y siendo conocidos, no son lo suficientemente fiables como para tenerlos en cuenta en el estudio. En estas circunstancias, siendo por cierto muy comunes cuando se lleva a cabo una investigación que requiera de las técnicas estadísticas multivariantes, hay varias formas de actuación que se resumen en lo sucesivo.

Es más, a la hora de resolver el problema propuesto, de la clasificación u ordenación de individuos a partir de un conjunto de datos multivariantes, se deben tener en cuenta distintos aspectos en los que conviene tener precauciones acerca de dichos datos proporcionados. En general, por su propia naturaleza, los datos pueden estar perturbados o pueden perturbar los resultados finales; entre otras, se deben observar las siguientes características:

- Independencia entre los casos (o individuos).
- Independencia entre las variables.
- Existencia de la totalidad de los valores a_{ij}.
- Definición correcta (y fiabilidad) de la totalidad de los a_{ij}.

Como se decía antes, cuando se verifiquen simultáneamente los cuatro puntos anteriores no debe existir problema alguno para analizar los datos por los procedimientos habituales. Por el contrario, a veces hay que utilizar metodologías diseñadas *ad hoc*. Aquí, nos centraremos en estudiar qué ocurre si no se verifica el punto tercero o el cuarto de la lista anterior.

Además, trataremos de ajustarnos a las condiciones más habituales en la práctica. En el problema de clasificación u ordenación propuesto, supongamos que los N individuos son indepedientes entre sí, pero que las M variables no son independientes; supondremos también que las variables (relacionadas) presentan valores

faltantes y que este problema afecta al análisis que se quiere realizar. Para resolver el problema, tradicionalmente se recomienda atender al número de variables y a los datos efectivamente conocidos de cada individuo, porque dependiendo de estos datos se puede utilizar una técnica u otra. En general, los datos perdidos se pueden clasificar como:

- Datos perdidos al azar (*missing at random*, MAR): los valores perdidos a_{ij} están asociados a variables Y_j; es decir, la probabilidad de que un valor no sea observado depende de los valores observados pero no de los valores faltantes.
- Datos perdidos completamente al azar (*missing completely at random*, MCAR): los valores perdidos a_{ij} no están relacionados con ningún valor de la variable ni individuos; es decir, el valor perdido no depende de los valores observados ni de los valores faltantes.
- Datos perdidos no al azar (*not missing at random*, NMAR): los valores a_{ij} están intrínsecamente relacionados con su propio valor; es decir, existe una dependencia entre los valores perdidos con los valores observados y valores faltantes.

A continuación se resumen algunas de las distintas técnicas existentes para el tratamientos de datos en presencia de valores perdidos:

3.1.3. Análisis de datos completos (*listwise* o *case deletion*, LD)

La eliminación de los datos incompletos es uno de los procedimientos más utilizados por la mayoría de los investigadores, en particular, en el ámbito de la

Ciencias Sociales. *Grosso modo*, la técnica consiste en eliminar del análisis el caso en el cual falte alguna observación.

Cuando se desea afrontar la clasificación de los distintos individuos mediante técnicas estadísticas determinísticas, en las cuales se debe conocer *a priori* toda la información, se requiere la eliminación de todos los casos a los que les falte alguna información. Este tratamiento resulta ser uno de los más restrictivos, al reducir el número de casos a N', perdiendo $N - N'$ individuos. La primera consecuencia obvia de la eliminación de casos es que con este primer tratamiento se reduce el número de datos analizados, por lo que también puede reducirse la información relevante.

Este procedimiento es aceptable siempre que al eliminar la información se verifiquen las dos siguientes condiciones:

- Los individuos eliminados tienen las mismas características que los datos completos existentes; es decir, no puede ocurrir que los individuos sin dato se comporten de manera esencialmente distinta que el resto del colectivo que se va a analizar.
- La falta de información se generó de manera aleatoria (un MCAR, según la notación anterior).

Sin embargo, en la mayoría de las situaciones empíricas no se verifica ninguna de estas dos hipótesis (sobre todo la primera), luego no sería conveniente aplicar esta estrategia. A pesar de eso, es una técnica muy utilizada en la práctica.

Lógicamente, en el caso del problema de clasificación, además de los posibles problemas de sesgo, no se podría clasificar a $N - N'$ individuos. Con esto, obtenemos la siguiente valoración del método:

- Ventajas: procedimiento muy simple.
- Inconvenientes: reducción de los datos, pues de N individuos se pasa a N'; pérdida de información; posible sesgo.

3.1.4. Análisis de datos disponibles (*pairwise deletion*, PD)

En el caso de observar los datos de manera transversal, atendiendo a las variables observadas, existen al menos dos formas distintas de tomar los datos disponibles. La primera forma se podría aplicar si de los N individuos se conocen M' variables de todos ellos. Entonces, se puede reducir el número de variables a M' y la muestra seguiría siendo de N individuos, pero despreciando $M - M'$ variables. Con este tratamiento, la clasificación de los N individuos estaría sesgada por la utilización de la información contenida exclusivamente en M' variables y no en las M que se deseaban incluir inicialmente.

Una variante para tratar los datos disponibles fue propuesta por Matthai [91] y consiste en realizar distintos análisis atendiendo a cada tipo de variable y no restringirse al número de variables que son conocidas para todos los individuos. Sin embargo, los resultados de los distintos análisis no serían comparables entre sí y no resolvería el problema de clasificación de manera global. En este caso no se reduce el número de variables, pero sí el número individuos utilizados para cada análisis (y su naturaleza), luego los análisis no son comparables y los individuos de los distintos análisis tampoco.

- Ventajas: procedimiento muy simple.
- Inconvenientes: reducción del número de variables o redución de los análisis comparables; posible sesgo.

3.1.5. Análisis con imputación de datos

Desde la propuesta de Rubin [125] en el año 1978, la imputación de datos es la alternativa llevada a cabo por la mayoría de los autores con conocimientos de Estadística, frente a las dos soluciones anteriores, si bien a lo largo de los años han surgido algunas modificaciones sobre el método original. Dependiendo de la proporción de los datos perdidos sobre el total, el uso de esta técnica puede provocar la perturbación de los datos y de los consiguientes resultados, ya que puede provocar subestimación de la verdadera varianza, tal y como afirman Rao y Shao [118].

Existen distintos modos de imputación de los datos perdidos, pero en todos ellos el fundamento es sustituir las observaciones faltantes por las obtenidas a partir de la información existente, mediante algún proceso de estimación estadística. Con este método se puede corregir el error de reducir el número de individuos y el número de variables, pero en cambio aumenta la subestimación de la verdadera varianza, como ya se ha apuntado.

En general, se puede clasificar la técnica de imputación en dos subtipos, como simple o múltiple. A continuación se resumen algunos de los métodos de imputación existentes.

- **Imputación simple:** las técnicas simples de imputación han sido y todavía son una herramienta muy utilizada y sencilla, aunque lógicamente menos eficaz que la imputación múltiple [126]. Para poder aplicar este tipo de imputación, los datos faltantes deben ajustarse al tipo MCAR.

  La imputación simple puede ser, a su vez, aleatoria o determinística. Tanto un tipo de imputación como el otro tienen sus ventajas y sus inconvenientes,

como puede ser que la imputación simple aleatoria tiene una mayor variablidad respecto a la imputación simple determinística, pero la imputación simple determinística suele ser más precisa que las correspondientes técnicas aleatorias. Algunas de las variantes más destacadas de las técnicas de imputación simple son:

- Técnicas de imputación simple aleatoria: los datos faltantes se actualizan con los datos conocidos, tomándolos como datos aleatorios.
- Técnicas de imputación simple determinística: son métodos directos, que consisten en la sustitución de los datos perdidos por la media, la mediana o la moda calculadas a partir de los valores conocidos. Este procedimiento afecta a la distribución, a su varianza, covarianza, quantiles, sesgo, etc.
- Imputación mediante la regresión: se estiman los valores perdidos utilizando la relación que existe entre los valores conocidos.
- Imputación mediante una RNA: se suele llevar a cabo una técnica similar a la de la imputación mediante regresión, calculando a partir de los datos conocidos, pero interviniendo una RNA en el proceso.
- Imputación por la función de máxima verosimilitud: se realizan las predicciones de los valores perdidos con el modelo que teóricamente mejor lo aproxima, utilizando para estimar los parámetros de este modelo los datos completos con su función de máxima verosimilitud. Estas predicciones se van actualizando cada vez que se calcula un valor perdido nuevo.

- **Imputación multiple:** a finales de los años 70 del siglo XX Rubin [125] propuso una imputación alternativa, la imputación múltiple. A la hora de

aplicar una imputación múltiple, se deben construir k conjuntos de datos completos ($k \geq 2$) y, utilizando el método de Monte Carlo, se sustituye cada dato faltante por los k valores obtenidos.

En todos estos tratamientos de imputación se comete un error (normalmente medido por el error cuadrático medio, ECM), ya que la estimación se realiza a partir de los datos de los individuos de los cuales se conocen todos los datos, pero estos individuos son supuestamente independientes unos de otros y se relacionan con variables con un grado de dependencia.

- Ventajas: utilización de todos los datos conocidos.
- Inconvenientes: imputar datos no conocidos a partir de los conocidos; posible sesgo.

3.1.6. Subsanación mediante RNA

El objetivo de este apartado es proponer un nuevo tratamiento para realizar una clasificación de los casos pertenecientes a una base de datos con valores faltantes, en la circunstancia de no poder aplicar o no ser adecuado ninguno de los tratamientos anteriores. Nuestra técnica se basa, según se verá, en la incorporación de las RNA, herramienta que ya ha sido utilizada por varios autores para imputar los valores perdidos o para aplicar el método de *network reduction* (NR), como se puede ver en [13], [141] y [12]; también se pueden encontrar referencias a algunas otras aplicaciones en [114], [142] y [132].

Aunque su origen es totalmente distinto, el método propuesto en esta tesis consiste en una variante del método NR, que inicialmente se utilizó en predicción.

En realidad, nosotros perseguimos una ordenación de un conjunto de individuos; dicha ordenación se obtendrá a partir de la generación de un indicador individual, pero para la obtención de dicho indicador se utilizarán RNA con el fin de establecer grupos que ayudarán al cálculo, como se explicará luego.

El método NR consiste en entrenar un conjunto de RNA para predecir la variable dependiente de un estudio, utilizando un vector de entrada distinto para cada una de las redes anteriores. El número de redes coincide con el número de grupos que a su vez viene determinado por los posibles patrones de datos faltantes que se puedan dar en el conjunto de variables. Además de para predecir, el modelo NR se puede utilizar para clasificar [132]. Autores como Boswell [13] afirman que esta técnica es más efectiva que la imputación aunque se utilice como una "imputación a partir de una red neuronal". Pero toda técnica tiene su inconveniente y en este caso la mayor es el gran número de redes de tipo *feedforward* (FF o PM con un entrenamiento específico, en la notación que estamos siguiendo) distintas que hay que entrenar convenientemente; este número viene determinado por los diferentes grupos que se deben analizar. En general, la consecuencia práctica es que se necesita determinar el valor de un gran número de parámetros.

La aportación que se sugiere a continuación es una variante de la técnica NR que también viene relacionada con el número de grupos y con la utilización de una RNA para clasificar. La principal diferencia es que no se necesita de una supervisión de los datos de salida *a priori*, mientras que en el método NR se necesita conocer el vector de salida para llevar acabo la clasificación o la predicción de los datos. Es decir, en nuestro caso la clasificación se realiza a partir de una RNA no supervisada. Con esta modificación se reduce la influencia del investigador y el número de parámetros a estimar, algo a tener muy en cuenta en la aplicación práctica, pues una red tipo FF incorpora como mínimo un número de parámetros

que se calcula como la longitud del vector de entrada más 1 más la longitud del vector de salida; y eso siempre y cuando no haya ninguna neurona en capa oculta. En nuestro caso, los parámetros que se actualizan y se calculan vienen determinados por el número de grupos que se desean clasificar. Por ejemplo, si se desea que la RNA clasifique el grupo de individuos en dos subgrupos, los parámetros a calcular y actualizar vienen determinados por los subgrupos calculados, lo que en este ejemplo serían 2. Para determinar el número de subgrupos más adecuado, en cada caso, para permitir la posterior clasificación se utilizará una cota que se definirá enseguida.

Descripción del procedimiento

Según se ha anticipado, proponemos un tratamiento original para obtener una puntuación para cada individuo. Dicha puntuación servirá para ordenar dichos individuos o para generar grupos o clases; bajo determinadas condiciones, también puede ser útil para comparar unos individuos con otros en estudios que persigan el análisis de una evolución o mejora.

Supongamos un grupo de N individuos; de N' de ellos se conocen los valores de las M variables del estudio, mientras que de los $N-N'$ restantes no se conocen todas las variables.

Entre los $N - N'$ individuos "incompletos", se pueden establecer grupos uniformes en cuanto a las variables de las que se dispone de datos. Llamemos p al número de grupos distintos y G_k a cada uno de los grupos. El número máximo de grupos será $2^M - 1$, luego $p \leq 2^M - 1$. Lógicamente, todos esos grupos son disjuntos dos a dos, ya que cada uno de los grupos viene definido por el conjunto de variables conocidas para todos los individuos de dicho grupo. Así, el conjun-

to de los N individuos se puede escribir como la unión disjunta $\cup_{k=1}^{p} G_k$, donde $G_i \cap G_j = \emptyset \, \forall i \neq j$. Denotaremos por N_k al cardinal de G_k. Así, es obvio que $\sum_{1}^{p} N_k = N$.

Una vez creados los p grupos y comprobado que están bien definidos (esto es, que todo individuo X_i pertenece a un único G_k, con $k = 1, ..., p$), se diseñan y entrenan p RNA, una por cada grupo G_k; las llamaremos RNA_k, con $k = 1, ..., p$. Todas ellas serán no supervisadas.

Cada grupo G_k viene caracterizado por el número de variables de las que se dispone de información para todos los individuos, r_k, así como por el número de elementos de dicho grupo, N_k. Por eso, cada RNA_k (con $k = 1, ..., p$) tiene por vector de datos de entrada los $X_i \in G_k \subset \mathbb{R}^{r_k}$ cuya longitud es, obviamente, r_k lo que coincide con el número de variables "correctamente definidas" en dicho grupo.

Para poder establecer una parada en el entrenamiento de la red no supervisada, definimos una "cota de parada" (ϑ) distinta para cada $RNA_k, k = 1, ..., p$:

Definición 3.1.1. *Dado el vector de entrada t_k, con longitud r_k, de la red no supervisada RNA_k, se define la cota de parada ϑ de la RNA_k como:*

$$\vartheta_k = \sqrt{\frac{r_k}{2}},$$

con $\vartheta_k \in (0, +\infty)$.

Utilizaremos la cota de la Definición 3.1.1 para parar el proceso de entrenamiento de la red si está suficientemente entrenada o si hace falta aumentar el número de subgrupos de la clasificación. Es decir, una vez entrenada la RNA_k con h_k subgrupos, el programa utilizado nos devuelve el ECM calculado como la distancia entre los vectores base y los casos utilizados para validar el entrenamiento. Así, asumiremos que el número adecuado de subgrupos es este h_k si:

$$\vartheta_k \geq ECM(RNA_k) \tag{3.1}$$

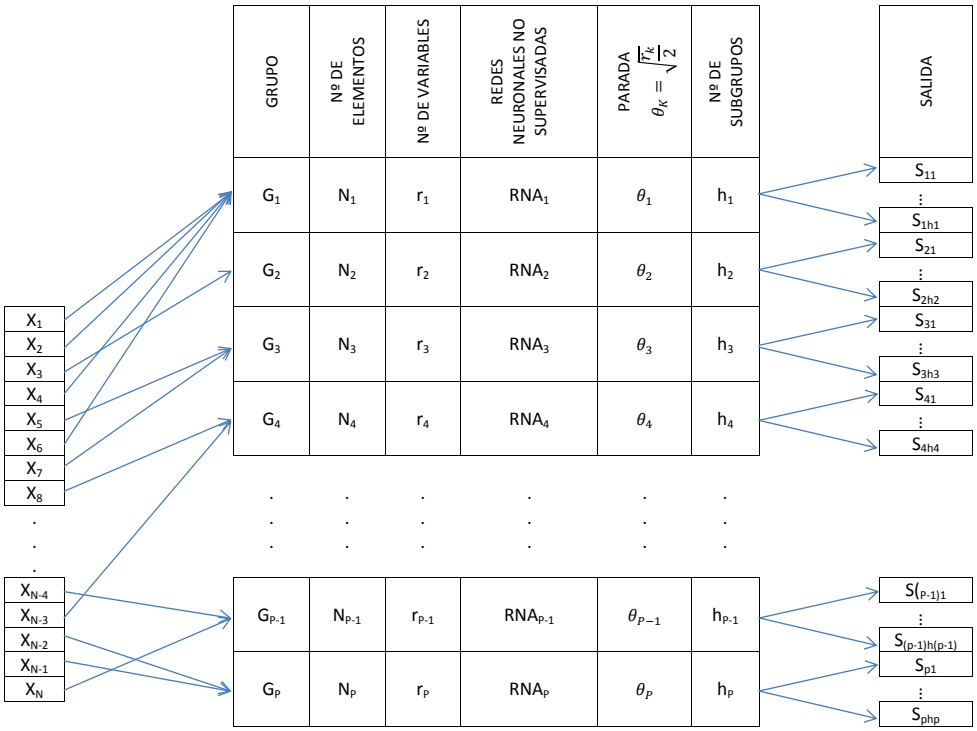
En el caso de no verificarse la Ecuación 3.1 para un cierto número de subgrupos h', se volverá a realizar el entrenamiento de la red RNA_k modificando el número de subgrupos de h' a $h'+1$.

Repitiendo el procedimiento anterior para cada G_k, se puede llegar a clasificar a todos los X_i, en $S \geq p$ subgrupos, siendo $S = \sum_{k=1}^{p} h_k$. En resumen, una vez entrenadas las redes, el resultado obtenido proporcionaría una primera clasificación de los individuos comparables (ver la Figura 3.1). Se habría generado un orden parcial útil para establecer posteriormente una puntuación a cada individuo. Y ello se ha logrado utilizando todos los datos propuestos, sin modificar ni añadir ningún valor y reduciendo los parámetros a calcular con respecto al método NR propuesto por otros autores.

- Ventajas: utilización de todos los datos conocidos; comparaciones entre semejantes, reduciendo el sesgo.

- Inconvenientes: el número de subgrupos puede ser muy elevado, pues se puede partir de $2^M - 1$ grupos que se elevarían a la hora de clasificar; cierta subjetividad al proponer la cota de parada.

Figura 3.1: Ilustración de una clasificación mediante RNA por subgrupos



Ejemplo

Supongamos que se cuenta con el siguiente conjunto de datos, en el que su información se recoje mediante la siguiente matriz de datos multivariantes:

$$A = \begin{pmatrix}
1 & 7 & 5 & - \\
2 & 3 & - & - \\
- & 3 & 4 & - \\
1 & 10 & 3 & 2 \\
1 & - & - & - \\
2 & 3 & 4 & 5 \\
- & - & 8 & 9 \\
7 & 8 & 9 & 4 \\
9 & 8 & 1 & 4 \\
- & 3 & - & - \\
- & 7 & 5 & - \\
7 & 3 & - & - \\
- & 5 & 5 & - \\
1 & 10 & 3 & 2 \\
1 & - & 8 & - \\
2 & 3 & 4 & 5 \\
- & - & 8 & 9 \\
7 & 8 & 9 & 4 \\
9 & 8 & 1 & 4 \\
- & 3 & 9 & - \\
- & 7 & 5 & - \\
2 & 3 & - & - \\
- & 3 & 5 & - \\
1 & 10 & 3 & 2 \\
8 & - & - & - \\
2 & 3 & 4 & 5 \\
- & - & 8 & 9 \\
7 & - & 9 & 4 \\
9 & 8 & 1 & 4 \\
- & 10 & - & -
\end{pmatrix}$$

Atendiendo a las dimensiones de la matriz, se puede observar que existe un total de 30 filas (es decir $N = 30$ casos) y el número de columnas viene determi-

nado por el número de variables, que en este caso es $M = 4$. Luego, por ejemplo, el vector $(1, 7, 5, -)$ corresponde a las 4 observaciones del individuo X_1 (la cuarta, a_{14}, sería un dato faltante). Observando la información existente, se puede apreciar que existen bastantes valores perdidos, aunque no conozcamos el motivo de su pérdida.

Para poder llevar acabo cualquier análisis tradicional necesitaríamos información de todas las variables e individuos. Por eso, se decide aplicar alguna técnica para el tratamiento de datos incompletos. Si utilizáramos exclusivamente los individuos con datos completos, nos quedaríamos con 11 casos de los 30 posibles; si se utilizara el método de variables completas, nos quedaríamos sin ninguna variable (en este caso, las cuatro variables presentan algún problema). Por ello, se decidió utilizar alguna técnica menos restrictiva, como la que acabamos de presentar, utilizando RNA.

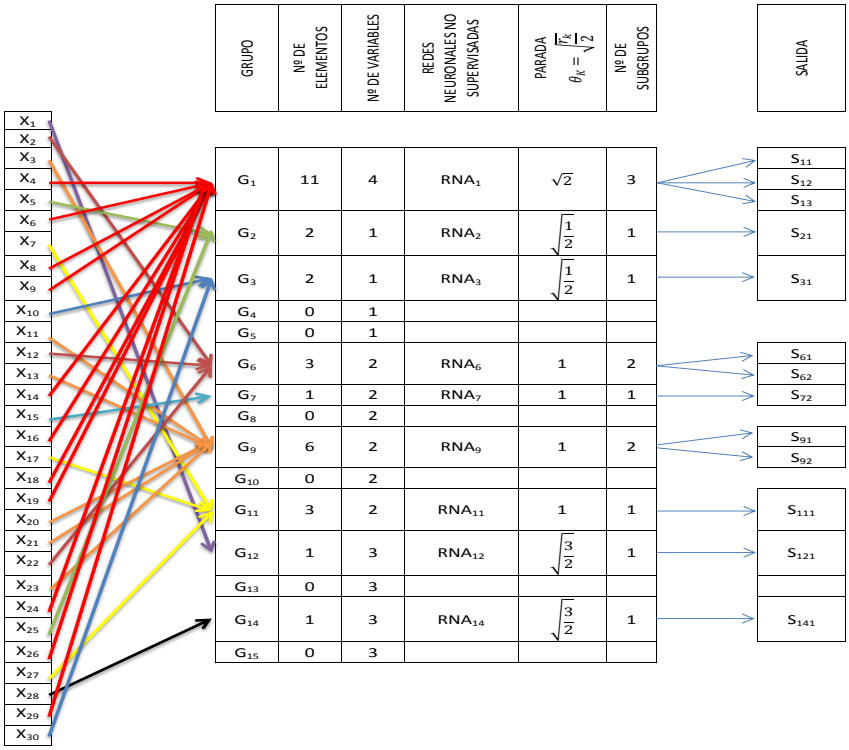
Para poder aplicar nuestra técnica de subsanación, en primer lugar se tienen que crear subgrupos disjuntos de casos. En este caso, el número de grupos distintos definidos (p en la notación anterior) sería 15 como máximo, pero algunos de ellos serían grupos sin elementos; en concreto, en este caso hay 6 combinaciones de variables faltantes que no se da para ningún caso y esto hace que se reduzca el número de parámetros a estimar.

La Figura 3.2 puede ayudar a entender mejor el procedimiento. En ella se puede observar, por ejemplo, para cada subgrupo: el número de casos (*elementos*), la dimensión de los vectores (*variables*) y la cota de error (*parada*) establecida. Tras entrenar las RNA no supervisadas, se obtiene un total de 13 subgrupos a partir de los 30 casos iniciales.

En este ejemplo suponemos que se puede utilizar una ordenación en cada uno de los subgrupos obtenidos, utilizado el conjunto de variables conocidas en cada caso, y que también se puede obtener un valor para cada subgrupo. Esto

quiere decir que la metodología permite realizar una primera clasificación de los 30 individuos (todos los casos) en 13 subgrupos. Sin embargo, hay un siguiente paso a seguir y que consistiría en establecer una ordenación y acotación de los distintos subgrupos, pudiendo llegar a reducirse el número final de ellos. Creemos que la aplicación de esta técnica a un caso real (lo cual se hará en el Capítulo 4) permitirá describir con más claridad las posibilidades con las que se cuenta en esta última parte.

Figura 3.2: Ilustración de un ejemplo de clasificación mediante RNA por subgrupos



3.2. Determinación de la RNA más adecuada

Uno de los principales problemas a los que se enfrenta un investigador a la hora de aplicar las RNA (como técnica de minería de datos, para clasificar, ordenar, ajustar, simular, etc.) es determinar cuáles son las características más apropiadas para la RNA según sea el conjunto de datos disponible y lo que se quiere conseguir de ellos. Es decir, elegir y utilizar una RNA más adecuada puede reducir significativamente el error cometido y el tiempo de computación empleado para especificar los distintos parámetros necesarios para obtener la RNA entrenada que lleve a resolver el problema planteado.

Atendiendo a la Definición 2.4.21, un investigador tiene que tener en cuenta las distintas partes de la RNA para diseñar una que se ajuste a la situación que afronta. Una de las partes fundamentales es la estructura topológica subyacente, por lo que, consecuentemente, uno de los primeros pasos debe ser determinar el digrafo asociado al conjunto de datos y a la estrategia de resolución del problema. De hecho, según sea el digrafo subyacente elegido, para un mismo conjunto de datos, el error cometido tras el entrenamiento puede variar considerablemente.

A continuación vamos a construir un ejemplo para comprobar cómo de sensibles son las RNA a las variaciones en su estructura topológica subyacente.

Considérese un conjunto de datos organizado en 80 variables de entrada y 5 variables de salida, con un conjunto total de 17 casos o individuos. Por las características comentadas de los datos, se puede inferir que hay exceso de variables para pocos individuos; es decir, de cada uno de los 17 casos hay información correspondiente a 80 variables y todos ellos se clasifican según la variable de salida, de dimensión 5 (por su propia naturaleza, podemos suponer que se trata de una clasificación de los individuos en 5 subgrupos disjuntos). Una vez descrito el conjunto de datos, ¿sería posible conocer qué RNA sería más interesante utilizar en

este problema de clasificación?

Atendiendo a la información de que tenemos 5 variables de salida, la primera intuición sería construir una RNA con solo una capa oculta constituida por 5 neuronas. Así, se programó una RNA de tipo perceptrón multicapa, con una sola capa oculta y dicha capa compuesta por 5 neuronas. Se obtuvieron los resultados que se muestran en la primera fila de la Tabla 3.1. Se puede deducir que, con solo 17 datos y 80 variables, se ha conseguido obtener una RNA que clasifica los individuos en 5 grupos con un $74,48\,\%$ de acierto.

Tabla 3.1: Resultado de entrenar una RNA con neuronas en la capa oculta

Neuronas en capa oculta	Error	Tiempo (segundos)
5	0,255288	1,264
8	0,193498	2,262

Sin embargo, cualquier investigador que entrene dicha RNA podría plantearse si utilizar otra estructura topológica podría servir para reducir el error cometido. Por ello, utilizamos otra RNA de la misma tipología, pero con una pequeña modificación: en lugar de tener 5 neuronas en la capa oculta, decidimos incorporar 8 neuronas. Observando la última fila de la Tabla 3.1, se puede apreciar que el incremento del número de neuronas consigue reducir el error del modelo generado, pues en esta segunda ocasión se logra un $80,75\,\%$ de acierto en la clasificación, aunque también se produce aumento de casi un segundo adicional en el tiempo de computación.

Es decir, en nuestro ejemplo, con una pequeña modificación en la estructura topológica se ha reducido el error; en este caso, parece obvio que el incremento de un segundo en el proceso computacional se debe al aumento del número de

parámetros a estimar, que es lógicamente mayor cuando se utiliza una red con 8 neuronas en la capa oculta que cuando solo había 5.

Ahora bien, los resultados de la Tabla 3.1 nos pueden llevar a preguntarnos si será realmente la RNA de 8 neuronas la que menos error cometa para este caso o si habrá otra más efectiva (y la pregunta podría complicarse si consideráramos la eficiencia, en cuanto a tiempo de cálculo). Pues esta cuestión es uno de los grandes retos en el campo de las RNA, ya que no es sencillo determinar la mejor estructura topológica *a priori* y tampoco es posible plantear todas las opciones posibles y analizarlas una a una (por las limitaciones de tiempo y memoria operativa, principalmente). Por todo esto, nos planteamos si es posible diseñar una RNA que sea capaz de ayudar a entrenar RNA. Antes de abordar este reto tan ambicioso, nos dedicamos a realizar un programa que sirviera de ayuda para valorar diferentes estructuras topológicas subyacentes, entrenando diferentes RNA y devolviendo la más adecuada para cada caso particular; esto es, se trataba de realizar un programa que devuelva qué RNA interesa más según la información con la que se cuente.

Por otra parte, si echamos un vistazo a la Definición 2.4.21, hay otros aspectos que también se deberían tener en cuenta a la hora de diseñar una RNA, aunque dependerán de la estructura topológica elegida; por ejemplo, las funciones neuronales asociadas a dicha estructura topológica pueden ser de varios tipos y, dependiendo de la salida que se desee obtener (es decir, una salida lineal o no lineal), se puede de utilizar una o otra.

Y, sobre la regla de aprendizaje, también se debe considerar que habitualmente las RNA dividen los datos en dos subconjuntos: el de los datos de validación y el de los datos de entrenamiento. Dependiendo de la división del conjunto de datos que se tome, se puede llegar a entrenar perfectamente la RNA o, por el contrario, se puede cometer un error considerable; en concreto, es importante elegir el conjunto

de datos de entrenamiento (y el conjunto de datos de validación) de modo que sea representativo del conjunto total de casos. Puesto que la representatividad suele ser una característica que se evalúa *a posteriori*, consideramos recomendable probar diferentes divisiones del conjunto de datos para obtener un entrenamiento lo más eficiente que sea posible.

Con las distintas ideas que acabamos de comentar, se motiva el desarrollo de un programa informático (implementado en el paquete de computación simbólica Mathematica y en su propio lenguaje de programación) que nos ayudará a detectar la RNA que mejor resuelva el problema planteado para el correspondiente conjunto de datos.

3.2.1. Implementación en Mathematica

En esta sección se presenta la implementación en Mathematica 9 de un programa que posibilita la evaluación de diferentes variantes de RNA útiles para resolver un mismo problema. Tal y como se ha comentado en el apartado anterior, este programa se ha desarrollado como una herramienta para detectar qué tipo de red (qué tipo de estructura topológica, qué forma de subdividir el conjunto de datos, qué elección para las funciones de activación, etc.) es la más adecuada para generar un modelo que aproxime la configuración de los datos. Aunque la innovación que representa no sea muy llamativa, resulta ser una herramienta con un coste computacional algo inferior a la realización de cada uno de los modelos por separado, además de que reduce el error calculado al mínimo entre las redes propuestas, sin necesitar de la intervención o supervisión continua por parte del investigador.

Inicialización

```
Check[
  ClearAll["Global`*"];
  << NeuralNetworks`;
  , Quit;
  ];

Check[
  ventanaSalida =
   CreateDocument[{}, WindowSize -> Scaled[1], WindowSelected -> True];
  NotebookWrite[ventanaSalida, Cell["Inicializando programa...", "Text"]];
  , Quit;
  ];

nombreProblema = "test1";
```

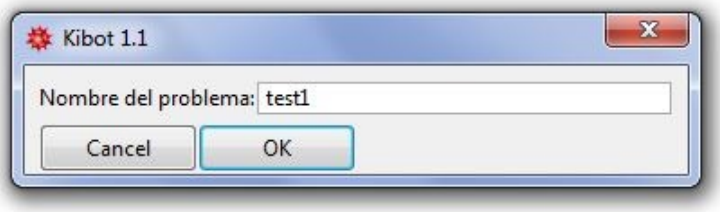


Figura 3.3: Pantalla de inicio del programa propuesto en Mathematica 9

Definición de constantes

```
tpNeuronasCapa1Desde = 1;
tpNeuronasCapa1Hasta = 2;
tpNeuronasCapa1Intervalo = 3;
tpNeuronasCapa2Desde = 4;
tpNeuronasCapa2Hasta = 5;
tpNeuronasCapa2Intervalo = 6;
tpMetodosParticion = 7;
tpInicializacionesNoAleatorias = 8;
tpInicializacionesAleatorias = 9;
tpFuncionSalidaNoLineal = 10;
tpIteracionesIntervalo = 11;
tpFactorFinalizacionEntrenamiento = 12;
trId = 1;
trTipo = 2;
trNeuronas = 3;
trMetodoParticion = 4;
trInicializacionAleatoria = 5;
trIteraciones = 6;
trMejorIteracion = 7;
trEstado = 8;
trLogErrorConjuntoEntrenamiento = 9;
trLogErrorConjuntoValidacion = 10;
trPctErrorConjuntoEntrenamiento = 11;
trPctErrorConjuntoValidacion = 12;
trPctErrorTotal = 13;
metodosParticion = {(*"1e",*)"3eev", "3eve", "3vee"(*,"3rot"*)};
funcionesSalidaNoLineal = {None, Sigmoid};
```

Lectura de archivos y parámetros

```
If[
  DialogInput[
    Column[{
      Row[{"Nombre del problema:",
        InputField[Dynamic[nombreProblema], String]}],
      Row[{CancelButton[DialogReturn[False]],
        DefaultButton[DialogReturn[True]]}]
      }],
    WindowTitle -> "Kibot 1.1"
    ] == False
  ,
  NotebookWrite[ventanaSalida,
   Cell["Operación cancelada por el usuario.", "Text"]];
  Quit[];
  ];

Check[
  SetDirectory[NotebookDirectory[] <> nombreProblema <> "\\"];
  datosEntrada = << entrada.dat;
  datosSalida = << salida.dat;
  NotebookWrite[ventanaSalida,
   Cell["Archivos de entrada y salida leídos.", "Text"]];
  , Quit[];
  ];
If[
  FileExistsQ["parametros.dat"]
  ,
  Check[
    parametros = << parametros.dat;
    NotebookWrite[ventanaSalida, Cell["Archivo de parámetros leído.",
```

```
    "Text"]];
    , Quit[]
    ];
  ,
  NotebookWrite[ventanaSalida,
   Cell["No hay archivo de parámetros; se generará uno nuevo.", "Text"]];
  parametros = Normal[SparseArray[{
      tpNeuronasCapa1Desde -> 0,
      tpNeuronasCapa1Hasta -> 20,
      tpNeuronasCapa1Intervalo -> 5,
      tpNeuronasCapa2Desde -> 0,
      tpNeuronasCapa2Hasta -> 10,
      tpNeuronasCapa2Intervalo -> 5,
      tpMetodosParticion -> Null,
      tpInicializacionesNoAleatorias -> 1,
      tpInicializacionesAleatorias -> 1,
      tpFuncionSalidaNoLineal -> funcionesSalidaNoLineal[[1]],
      tpIteracionesIntervalo -> 30,
      tpFactorFinalizacionEntrenamiento -> 2
      }]];
  parametros[[tpMetodosParticion]] = {metodosParticion[[1]]};
  ];

a = parametros[[tpMetodosParticion]];
If[
  DialogInput[
    Column[{
      Row[{"Nombre del problema: ", nombreProblema}],
      Row[{"Dimensión de los datos de entrada: ",
        If[Length[Dimensions[datosEntrada]] > 1,
         Dimensions[datosEntrada][[2]], 1]}],
      Row[{"Dimensión de los datos de salida: ",
```

```
  If[Length[Dimensions[datosSalida]] > 1, Dimensions[datosSalida]
  [[2]],1]}],
Row[{"Número de pares de datos de entrada/salida: ",
  Dimensions[datosEntrada][[1]]}],
Row[{"Neuronas capa 1: desde ",
  InputField[Dynamic[parametros[[tpNeuronasCapa1Desde]]], Number,
   FieldSize -> 5],
  " hasta ",
  InputField[Dynamic[parametros[[tpNeuronasCapa1Hasta]]], Number,
   FieldSize -> 5],
  " intervalo ",
  InputField[Dynamic[parametros[[tpNeuronasCapa1Intervalo]]], Number,
   FieldSize -> 5]}],
Row[{"Neuronas capa 2: desde ",
  InputField[Dynamic[parametros[[tpNeuronasCapa2Desde]]], Number,
   FieldSize -> 5],
  " hasta ",
  InputField[Dynamic[parametros[[tpNeuronasCapa2Hasta]]], Number,
   FieldSize -> 5],
  " intervalo ",
  InputField[Dynamic[parametros[[tpNeuronasCapa2Intervalo]]], Number,
   FieldSize -> 5]}],
Row[{"Métodos de partición: ",
  CheckboxBar[Dynamic[a], metodosParticion]}],
Row[{"Función de salida no lineal: ",
  RadioButtonBar[Dynamic[parametros[[tpFuncionSalidaNoLineal]]],
   funcionesSalidaNoLineal]}],
Row[{"Número de inicializaciones:",
  " no aleatorias: ",
  InputField[Dynamic[parametros[[tpInicializacionesNoAleatorias]]],
   Number, FieldSize -> 3],
  " aleatorias: ",
```

```
        InputField[Dynamic[parametros[[tpInicializacionesAleatorias]]],
         Number, FieldSize -> 3]}],
      Row[{"Número de iteraciones por intervalo de entrenamiento: ",
        InputField[Dynamic[parametros[[tpIteracionesIntervalo]]], Number,
         FieldSize -> 3]}],
      Row[{"Factor de finalización de entrenamiento: ",
        InputField[Dynamic[parametros[[tpFactorFinalizacionEntrenamiento]]],
         Number, FieldSize -> 3]}],
      Row[{CancelButton[DialogReturn[False]],
        DefaultButton[DialogReturn[True]]}]
      }],
    WindowTitle -> "Kibot 1.1"
    ] == False
  ,
  NotebookWrite[ventanaSalida,
   Cell["Operación cancelada por el usuario.", "Text"]];
  Quit[];
  ];
parametros[[tpMetodosParticion]] = a;
Clear[a];

Check[
  parametros >> parametros.dat;
  NotebookWrite[ventanaSalida,
   Cell["Archivo de parámetros guardado.",
   "Text"]];
  , Quit[]
  ];

If[
  FileExistsQ["redes.dat"],
  Check[
```

```
    redes = << redes.dat
    , Quit[]
    ];,
  redes = {};
  NotebookWrite[ventanaSalida,
   Cell["No hay archivo de redes; se generará uno nuevo.", "Text"]];
  ];

Do[
  If[True
     && (i > 0 || j == 0)
     && Length[
       Select[redes, True
          && Length[#] >= trNeuronas
          && #[[trNeuronas]] == {i, j}
          && #[[trMetodoParticion]] == parametros[[tpMetodosParticion, k]]
          && #[[
            trInicializacionAleatoria]] == (l >
             parametros[[tpInicializacionesNoAleatorias]])
         &]
       ] <
      If[l > parametros[[tpInicializacionesNoAleatorias]],
       parametros[[tpInicializacionesAleatorias]],
       parametros[[tpInicializacionesNoAleatorias]]],
    AppendTo[redes, Normal[SparseArray[{
        trId -> Length[redes] + 1,
        trTipo -> "FF",
        trNeuronas -> Null,
        trMetodoParticion -> parametros[[tpMetodosParticion, k]],
        trInicializacionAleatoria -> (l >
           parametros[[tpInicializacionesNoAleatorias]]),
        trIteraciones -> 0,
```

```
        trMejorIteracion -> 0,
        trEstado -> Null,
        trLogErrorConjuntoEntrenamiento -> Null,
        trLogErrorConjuntoValidacion -> Null,
        trPctErrorConjuntoEntrenamiento -> Null,
        trPctErrorConjuntoValidacion -> Null,
        trPctErrorTotal -> Null
        }] ]];
    redes[[Length[redes], trNeuronas]] = {i, j};
    ];
  , {j, parametros[[tpNeuronasCapa2Desde]],
   parametros[[tpNeuronasCapa2Hasta]], parametros[[tpNeuronasCapa2Intervalo]]}
  , {i, parametros[[tpNeuronasCapa1Desde]],
   parametros[[tpNeuronasCapa1Hasta]], parametros[[tpNeuronasCapa1Intervalo]]}
  , {k, 1, Length[parametros[[tpMetodosParticion]]]}
  , {l, 1,
   parametros[[tpInicializacionesNoAleatorias]] +
    parametros[[tpInicializacionesAleatorias]]}
  ];
Clear[i, j, k, l];

Check[
  redes >> redes.dat;
  NotebookWrite[ventanaSalida, Cell["Archivo de redes guardado.", "Text"]];
  , Quit[]
  ];
```

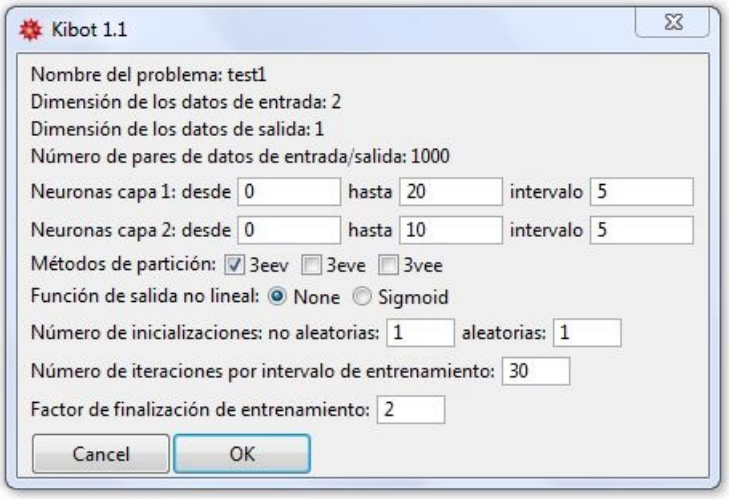


Figura 3.4: Definición de parámetros

Mostrar información

```
NotebookWrite[ventanaSalida, Cell[BoxData[ToBoxes[Grid[{
       {"PARÁMETROS", SpanFromLeft},
       {"Nombre del problema", nombreProblema},
       {"Dimensión de los datos de entrada",
        ToString[If[Length[Dimensions[datosEntrada]] > 1,
          Dimensions[datosEntrada][[2]], 1]]},
       {"Dimensión de los datos de salida",
        ToString[If[Length[Dimensions[datosSalida]] > 1,
          Dimensions[datosSalida][[2]], 1]]},
       {"Número de pares de datos de entrada/salida",
        ToString[Dimensions[datosEntrada][[1]]]},
       {"Neuronas capa 1",
        Table[i, {i, parametros[[tpNeuronasCapa1Desde]],
          parametros[[tpNeuronasCapa1Hasta]],
          parametros[[tpNeuronasCapa1Intervalo]]}]},
       {"Neuronas capa 2",
        Table[i, {i, parametros[[tpNeuronasCapa2Desde]],
          parametros[[tpNeuronasCapa2Hasta]],
          parametros[[tpNeuronasCapa2Intervalo]]}]},
       {"Métodos de partición", parametros[[tpMetodosParticion]]},
       {"Función de salida no lineal", parametros[[tpFuncionSalidaNoLineal]]},
       {"Número de inicializaciones no aleatorias",
        parametros[[tpInicializacionesNoAleatorias]]},
       {"Número de inicializaciones aleatorias",
        parametros[[tpInicializacionesAleatorias]]},
       {"Número de iteraciones por intervalo de entrenamiento",
        parametros[[tpIteracionesIntervalo]]},
       {"Factor de finalización de entrenamiento",
        parametros[[tpFactorFinalizacionEntrenamiento]]}
```

```
        }, Frame -> All]]], "Text", ShowStringCharacters -> False]];

NotebookWrite[ventanaSalida, Cell[BoxData[ToBoxes[Dynamic[Grid[Join[
        {{"TODAS LAS REDES", SpanFromLeft}},
        {{"Id", "Tipo", "Neuronas", "MetPart", "InicAleat", "Iter",
          "MejorIter", "Estado", "LogErrEnt", "LogErrVal", "%ErrEnt",
          "%ErrVal", "%ErrTot"}},
        redes
        ], Frame -> All]]]], "Text", ShowStringCharacters -> False]];

NotebookWrite[ventanaSalida, Cell[BoxData[ToBoxes[Dynamic[Grid[Join[
        {{"MEJORES REDES EN LOGARITMO-ERROR-VALIDACION", SpanFromLeft}},
        {{"Id", "Tipo", "Neuronas", "MetPart", "InicAleat", "Iter",
          "MejorIter", "Estado", "LogErrEnt", "LogErrVal", "%ErrEnt",
          "%ErrVal", "%ErrTot"}},
        SortBy[redes, #[[trLogErrorConjuntoValidacion]] &][[1 ;; 10]]
        ], Frame -> All]]]], "Text", ShowStringCharacters -> False]];

NotebookWrite[ventanaSalida, Cell[BoxData[ToBoxes[Dynamic[Grid[Join[
        {{"MEJORES REDES EN PORCENTAJE-ERROR-VALIDACION", SpanFromLeft}},
        {{"Id", "Tipo", "Neuronas", "MetPart", "InicAleat", "Iter",
          "MejorIter", "Estado", "LogErrEnt", "LogErrVal", "%ErrEnt",
          "%ErrVal", "%ErrTot"}},
        SortBy[redes, #[[trPctErrorConjuntoValidacion]] &][[1 ;; 10]]
        ], Frame -> All]]]], "Text", ShowStringCharacters -> False]];
```

Iteración

```
abortar = False;
NotebookWrite[ventanaSalida,
  Cell[BoxData[
    ToBoxes[Button["Abortar entrenamientos", abortar = True,
      Method -> "Preemptive", Enabled -> Dynamic[! abortar]]]]]];

estadoEntrenamientos = "";
NotebookWrite[ventanaSalida,
  Cell[BoxData[ToBoxes[Dynamic[estadoEntrenamientos]]], "Text"]];

While[! abortar,

  redesDisponibles = Select[redes, #[[trEstado]] == Null &];
  If[Length[redesDisponibles] == 0, Break[];];
  redesOrdenadas =
   Sort[redesDisponibles, (#1[[trIteraciones]] + #1[[trId]]/1000) < (#2[[
         trIteraciones]] + #2[[trId]]/1000) &];

  rId = redesOrdenadas[[1, trId]];
  r = Position[redes[[All, trId]], rId, 1, 1][[1, 1]];
  estadoEntrenamientos = "Red #" <> ToString[rId] <> ":";

  estadoEntrenamientos =
   estadoEntrenamientos <> " Preparando conjuntos de datos...";
  Switch[redes[[r, trMetodoParticion]]
   , "1e",
   xe = datosEntrada;
   xv = {};
   ye = datosSalida;
```

```
 yv = {};
 , "3eev",
 xe = Drop[datosEntrada, {3, -1, 3}];
 xv = Take[datosEntrada, {3, -1, 3}];
 ye = Drop[datosSalida, {3, -1, 3}];
 yv = Take[datosSalida, {3, -1, 3}];
 , "3eve",
 xe = Drop[datosEntrada, {2, -1, 3}];
 xv = Take[datosEntrada, {2, -1, 3}];
 ye = Drop[datosSalida, {2, -1, 3}];
 yv = Take[datosSalida, {2, -1, 3}];
 , "3vee",
 xe = Drop[datosEntrada, {1, -1, 3}];
 xv = Take[datosEntrada, {1, -1, 3}];
 ye = Drop[datosSalida, {1, -1, 3}];
 yv = Take[datosSalida, {1, -1, 3}];
 , _,
 Interrupt[];
 ];

rNombreArchivo = "red" <> StringTake["00" <> ToString[rId], -3] <> ".dat";
If[FileExistsQ[rNombreArchivo]
 ,
 estadoEntrenamientos =
  estadoEntrenamientos <> " Leyendo " <> rNombreArchivo <> "...";
 rRegistro = Get[rNombreArchivo];
 rRegistro[[1, 1]] =
  OptionValue[rRegistro[[2]], ParameterRecord][[
   Length[OptionValue[rRegistro[[2]], ParameterRecord]]]];
 ,
 estadoEntrenamientos = estadoEntrenamientos <> " Inicializando nueva red...";
 Switch[redes[[r, trTipo]]
```

```
  , "FF",
  Quiet[rRegistro =
      InitializeFeedForwardNet[xe, ye,
       Select[redes[[r, trNeuronas]], # > 0 &],
       RandomInitialization -> redes[[r, trInicializacionAleatoria]],
       OutputNonlinearity -> parametros[[tpFuncionSalidaNoLineal]]];];
  , -,
  Interrupt[];
  ];
 ];

estadoEntrenamientos = estadoEntrenamientos <> " Entrenando red...";
If[Length[xv] > 0
 ,
 Quiet[{Null, rRegistro} =
    NeuralFit[rRegistro, xe, ye, xv, yv,
     parametros[[tpIteracionesIntervalo]], Method -> LevenbergMarquardt,
     CriterionLog -> False, CriterionPlot -> False]];
 ,
 Interrupt[];
 Quiet[{Null, rRegistro} =
   NeuralFit[rRegistro, xe, ye, parametros[[tpIteracionesIntervalo]],
    Method -> LevenbergMarquardt, CriterionLog -> False,
    CriterionPlot -> False]];
 ];

estadoEntrenamientos =
 estadoEntrenamientos <> " Guardando " <> rNombreArchivo <> "...";
Check[Put[rRegistro, rNombreArchivo];, Quit[]];

estadoEntrenamientos = estadoEntrenamientos <> " Leyendo resultados...";
If[Length[OptionValue[rRegistro[[2]], ParameterRecord]] - 1 -
```

```
    redes[[r, trIteraciones]] < parametros[[tpIteracionesIntervalo]],
 redes[[r, trEstado]] = "Detenida";];
redes[[r, trIteraciones]] =
 Length[OptionValue[rRegistro[[2]], ParameterRecord]] - 1;
If[parametros[[tpFuncionSalidaNoLineal]] == Sigmoid,
 redes[[r, trPctErrorConjuntoEntrenamiento]] =
  Round[100*
    Sum[If[UnitStep[rRegistro[[1]][xe[[i]]] - 0.5] != ye[[i]], 1, 0], {i, 1,
        Length[xe]}]/Length[xe], 0.01];
 redes[[r, trPctErrorTotal]] =
  Round[100*
    Sum[If[UnitStep[rRegistro[[1]][datosEntrada[[i]]] - 0.5] !=
        datosSalida[[i]], 1, 0], {i, 1, Length[datosEntrada]}]/
     Length[datosEntrada], 0.01];
 ];
If[Length[xv] > 0
 ,
 redes[[r, trMejorIteracion]] =
  Ordering[OptionValue[rRegistro[[2]], CriterionValidationValues], 1][[
    1]] - 1;
 If[redes[[r, trEstado]] == Null &&
   redes[[r, trMejorIteracion]]*
     parametros[[tpFactorFinalizacionEntrenamiento]] <
    redes[[r, trIteraciones]], redes[[r, trEstado]] = "Entrenada";];
 redes[[r, trLogErrorConjuntoEntrenamiento]] =
  Round[Log[10,
    OptionValue[rRegistro[[2]], CriterionValues][[
     redes[[r, trMejorIteracion]] + 1]]], 0.1];
 redes[[r, trLogErrorConjuntoValidacion]] =
  Round[Log[10,
    OptionValue[rRegistro[[2]], CriterionValidationValues][[
     redes[[r, trMejorIteracion]] + 1]]], 0.1];
```

```
   If[parametros[[tpFuncionSalidaNoLineal]] == Sigmoid,
    redes[[r, trPctErrorConjuntoValidacion]] =
      Round[100*
        Sum[If[UnitStep[rRegistro[[1]][xv[[i]]] - 0.5] != yv[[i]], 1, 0], {i,
           1, Length[xv]}]/Length[xv], 0.01];
    ];
   ,
   redes[[r, trLogErrorConjuntoEntrenamiento]] =
    Round[Log[10, Min[OptionValue[rRegistro[[2]], CriterionValues]]], 0.1];
   Interrupt[];
   ];

  estadoEntrenamientos = estadoEntrenamientos <> " Guardando redes.dat...";
  Check[redes >> redes.dat;, Quit[]];
  estadoEntrenamientos = estadoEntrenamientos <> " OK.";

  If[Length[Cells[ventanaSalida]] == 0, abortar = True];

  ];

If[abortar,
  NotebookWrite[ventanaSalida,
    Cell["Operación cancelada por el usuario.", "Text"]];,
  NotebookWrite[ventanaSalida, Cell["Entrenamientos finalizados.", "Text"]];
  ];

Clear[abortar, redesDisponibles, redesOrdenadas, r, rId, rNombreArchivo,
  rRegistro];
```

3.2.2. Ejemplo de búsqueda de la RNA más adecuada

A continuación se presenta, a modo de ilustración, un caso real en el cual se ha utilizado el programa anterior para elegir la RNA más adecuada a partir de los datos. Los datos utilizados en este ejemplo pueden consultarse en el Anexo A. A continuación se describen los pasos seguidos.

1. Se crean dos archivos de datos ".dat": uno con los valores de entrada y otro con los valores de salida. Los valores de entrada pueden estar definidos en forma binaria o continua (en este ejemplo los valores de entrada son valores discretos acotados entre 0 y 4); en cambio, los valores de salida tienen que estar definidos de forma binaria, para poder establecer la clasificación en los subgrupos deseados, que en este caso son 5.

2. Una vez creados ambos archivos, inicializamos el programa para la elección de la RNA.

3. Se deben definir ciertos parámetros para encontrar la RNA más adecuada; la búsqueda se verá acotada por los parámetros establecidos. En este ejemplo, los parámetros quedan definidos según se recoge en la Tabla 3.2.

4. Una vez leídos y actualizados los valores para los distintos parámetros, el programa se encarga de diseñar y entrenar las distintas RNA con las combinaciones posibles (según los parámetros).

5. Se obtienen las siguientes tablas con los resultados de la ejecución del programa. En concreto, los resultados se presentan mediante un conjunto de tres tablas, en las que se describen las siguientes características de cada una de las RNA consideradas.

   ID: número identificador de la RNA.

Tabla 3.2: Definición de los parámetros para buscar la RNA más adecuada

PARÁMETROS	
Nombre del problema	Ejemplo
Dimensión de los datos de entrada	80
Dimensión de los datos de salida	5
Número de pares de datos de entrada/salida	17
Neuronas capa 1	{0, 4, 8, 12, 16, 20}
Neuronas capa 2	{0, 2, 4, 6, 8, 10}
Métodos de partición	{3eev, 3eve, 3vee}
Función de salida no lineal	Sigmoid
Número de inicializaciones no aleatorias	1
Número de inicializaciones aleatorias	1
Número de iteraciones por intervalo de entrenamiento	30
Factor de finalización de entrenamiento	2

Fuente: elaboración propia

Tipo: tipología de la RNA aplicada.

Neuronas: devuelve un par de datos {a,b}, donde a corresponde al número de neuronas utilizadas en la primera capa oculta y b el número de neuronas utilizadas en la segunda capa oculta.

Método: se refiere a la forma de elección de los conjuntos de datos de entrenamiento y de datos de validación. Como los datos se dividen en $\frac{2}{3}$ partes para el entrenamiento y $\frac{1}{3}$ para la validación, dichos conjuntos se pueden tomar de tres formas distintas: “eev”, “eve” y “vee”.

Inicio: los valores de los pesos iniciales pueden ser aleatorios o no; aquí se especifica si esta inicialización se calcula aleatoriamente o si se utiliza un valor previo calculado de algún modo a partir de los datos.

Iter.: número de iteracciones efectuadas durante el entrenamiento.

Mejor iter.: devuelve el número de la iteracción donde se ha cometido menor error, es decir, la iteración en la que se contaba con una RNA

mejor entrenada en términos de error estimado.

Estado: el estado puede ser "Entrenada" o "Detenida" y este resultado viene determinado según si se ha conseguido entrenar la RNA o, por el contrario, si se ha detenido el entrenamiento sin culminar el entrenamiento (según los criterios habituales del paquete Neural Networks de Mathematica).

Log. error ent.: devuelve el logaritmo del error de entrenamiento (error sobre el conjunto de datos de entrenamiento) de las RNA.

Log. error val.: devuelve el logaritmo del error de validación (error sobre el conjunto de datos de validación) de las RNA.

% error ent.: devuelve el porcentaje de errores cometidos sobre el conjunto de entrenamiento, una vez tomada la RNA mejor entrenada.

% error val.: devuelve el porcentaje de errores cometidos sobre el conjunto de validación, una vez tomada la RNA mejor entrenada.

% error total: devuelve la media aritmética entre el porcentaje de error de validación y el porcentaje de error de entrenamiento.

En la Tabla 3.3 se presentan los resultados obtenidos al aplicar el programa sobre los datos comentados anteriormente (nótese que el programa utiliza el punto decimal anglosajón en lugar de la coma decimal española).

Como se puede observar en la Tabla 3.3, el programa nos devuelve el conjunto de todas las RNA entrenadas según los valores inicialmente designados para los parámetros; en este caso particular, se han entrenado 186 RNA. Si atendemos a la información que puede extraerse de dicha Tabla 3.3, se puede observar que algunas de las RNA se han entrenado perfectamente pero otras, en cambio, han visto detenidos sus entrenamientos. También creemos conveniente detenernos en el momento de la elección de la RNA más adecuada en cada caso, pues pueden existir diferentes RNA con porcentajes de error similares. Así, en el ejemplo, hemos encontrado varias redes con error de entrenamiento del 0 %, con lo que serían candidatas para ser seleccionadas, pero dicho error no quiere decir que se trate de la RNA más adecuada, ya que el porcentaje de error de validación es más interesante y puede ser mayor al 0 % en dichos casos. En resumen, hay RNA sin error de entrenamiento (o con error de entrenamiento bajo) que no son apropiadas porque pueden haberse "sobreentrenado" con los datos de partida y ese tipo de RNA puede no ser excesivamente útil, por lo que normalmente se debe despreciar.

Con esto se puede entender que resulta complicado valorar inmediatamente el comportamiento de todas las RNA entrenadas por nuestro programa (para así elegir la RNA más adecuada). A modo de ayuda para el investigador, el programa devuelve dos tablas bastante útiles: una con las 10 RNA con menor logaritmo del error de validación (Tabla 3.11) y otra con las 10 RNA con menor porcentaje del error de validación (Tabla 3.12).

En el ejemplo que estamos desarrollando, según la Tabla 3.12, la primera RNA que se presenta como adecuada es la RNA cuyo número de identificación es el 7. Dicha RNA tiene una estructura topológica realmente simple, con una sola capa oculta compuesta por 4 neuronas. En ese caso, el método de entrenamiento de los datos ha sido "eev", los valores de los pesos iniciales no han sido elegidos aleatoriamente y se han completado 61 iteraciones, llegando a encontrar el mejor modelo

Tabla 3.3: Resultados de los entrenamientos de las RNA del ejemplo de búsqueda

ID	Tipo	Neuronas	Método	Inicio	Iter.	Mejor iter.	Estado	Log. error ent.	Log. error val.	% error ent.	% error val.	% error total
1	FF	{0,0}	3eev	False	30	5	Entrenada	-1.1	-0.5	0	40	11.76
2	FF	{0,0}	3eev	True	61	18	Entrenada	-4.7	-0.5	0	40	11.76
3	FF	{0,0}	3eve	False	30	9	Entrenada	-2	-0.6	0	50	17.65
4	FF	{0,0}	3eve	True	61	27	Entrenada	-5.4	-0.6	0	50	17.65
5	FF	{0,0}	3vee	False	30	1	Entrenada	-0.5	-0.4	0	100	35.29
6	FF	{0,0}	3vee	True	30	6	Entrenada	-0.6	-0.4	36.36	83.33	52.94
7	FF	{4,0}	3eev	False	61	26	Entrenada	-1	-0.8	0	0	0
8	FF	{4,0}	3eev	True	92	33	Entrenada	-0.7	-0.5	16.67	20	17.65
9	FF	{4,0}	3eve	False	30	11	Entrenada	-0.6	-0.4	18.18	66.67	35.29
10	FF	{4,0}	3eve	True	92	37	Entrenada	-0.5	-0.5	63.64	66.67	64.71
11	FF	{4,0}	3vee	False	30	2	Entrenada	-0.6	-0.4	36.36	100	58.82
12	FF	{4,0}	3vee	True	30	4	Entrenada	-0.5	-0.4	72.73	83.33	76.47
13	FF	{8,0}	3eev	False	61	27	Entrenada	-3.8	-0.5	0	40	11.76
14	FF	{8,0}	3eev	True	30	8	Entrenada	-1	-0.4	0	40	11.76
15	FF	{8,0}	3eve	False	80	46	Detenida	-8.1	-0.5	0	50	17.65
16	FF	{8,0}	3eve	True	79	41	Detenida	-6.7	-0.7	0	16.67	5.88
17	FF	{8,0}	3vee	False	30	3	Entrenada	-0.9	-0.4	0	66.67	23.53
18	FF	{8,0}	3vee	True	30	7	Entrenada	-0.8	-0.4	0	83.33	29.41
19	FF	{12,0}	3eev	False	30	13	Entrenada	-2.1	-0.4	0	80	23.53
20	FF	{12,0}	3eev	True	61	27	Entrenada	-1.4	-0.6	0	40	11.76
21	FF	{12,0}	3eve	False	86	59	Detenida	-16.4	-0.4	0	50	17.65
22	FF	{12,0}	3eve	True	97	92	Detenida	-15.7	-0.7	0	33.33	11.76
23	FF	{12,0}	3vee	False	30	3	Entrenada	-0.8	-0.4	0	66.67	23.53
24	FF	{12,0}	3vee	True	30	5	Entrenada	-0.7	-0.4	0	66.67	23.53
25	FF	{16,0}	3eev	False	30	3	Entrenada	-0.9	-0.4	0	80	23.53

Tabla 3.4: Resultados de los entrenamientos de las RNA del ejemplo de búsqueda (parte 2)

ID	Tipo	Neuronas	Método	Inicio	Iter.	Mejor iter.	Estado	Log. error ent.	Log. error val.	% error ent.	% error val.	% error total
26	FF	{16,0}	3eev	True	113	81	Detenida	-7.1	-0.5	0	40	11.76
27	FF	{16,0}	3eve	False	30	12	Entrenada	-1.3	-0.4	0	100	35.29
28	FF	{16,0}	3eve	True	89	38	Detenida	-2.1	-0.7	0	16.67	5.88
29	FF	{16,0}	3vee	False	30	1	Entrenada	-0.6	-0.4	0	66.67	23.53
30	FF	{16,0}	3vee	True	30	3	Entrenada	-0.6	-0.3	18.18	83.33	41.18
31	FF	{20,0}	3eev	False	30	11	Entrenada	-1.4	-0.7	0	20	5.88
32	FF	{20,0}	3eev	True	61	15	Entrenada	-1	-0.5	0	80	23.53
33	FF	{20,0}	3eve	False	61	20	Entrenada	-1.7	-0.5	0	66.67	23.53
34	FF	{20,0}	3eve	True	92	42	Entrenada	-4.7	-0.6	0	33.33	11.76
35	FF	{20,0}	3vee	False	30	1	Entrenada	-0.6	-0.4	9.09	83.33	35.29
36	FF	{20,0}	3vee	True	30	4	Entrenada	-0.7	-0.4	9.09	66.67	29.41
37	FF	{4,2}	3eev	False	30	3	Entrenada	-0.4	-0.4	75	80	76.47
38	FF	4,2	3eev	True	91	41	Entrenada	-2.5	-0.6	0	20	5.88
39	FF	4,2	3eve	False	92	34	Entrenada	-0.6	-0.5	18.18	50	29.41
40	FF	4,2	3eve	True	185	86	Entrenada	-0.8	-0.6	0	50	17.65
41	FF	4,2	3vee	False	30	4	Entrenada	-0.5	-0.4	36.36	66.67	47.06
42	FF	4,2	3vee	True	61	20	Entrenada	-0.6	-0.4	9.09	50	23.53
43	FF	8,2	3eev	False	30	9	Entrenada	-0.6	-0.4	66.67	80	70.59
44	FF	8,2	3eev	True	61	30	Entrenada	-0.7	-0.7	25	20	23.53
45	FF	8,2	3eve	False	61	15	Entrenada	-0.6	-0.5	27.27	66.67	41.18
46	FF	8,2	3eve	True	30	5	Entrenada	-0.5	-0.4	100	100	100
47	FF	8,2	3vee	False	30	4	Entrenada	-0.5	-0.4	45.45	100	64.71
48	FF	8,2	3vee	True	30	11	Entrenada	-0.5	-0.4	72.73	100	82.35
49	FF	12,2	3eev	False	30	1	Entrenada	-0.4	-0.4	100	100	100
50	FF	12,2	3eev	True	30	13	Entrenada	-0.5	-0.4	33.33	80	47.06

Tabla 3.5: Resultados de los entrenamientos de las RNA del ejemplo de búsqueda (parte 3)

ID	Tipo	Neuronas	Método	Inicio	Iter.	Mejor iter.	Estado	Log. error ent.	Log. error val.	% error ent.	% error val.	% error total
51	FF	12,2	3eve	False	61	26	Entrenada	-1.1	-0.5	0	33.33	11.76
52	FF	12,2	3eve	True	109	55	Detenida	-1.9	-0.6	0	50	17.65
53	FF	12,2	3vee	False	30	1	Entrenada	-0.5	-0.4	90.91	100	94.12
54	FF	12,2	3vee	True	30	3	Entrenada	-0.5	-0.4	81.82	100	88.24
55	FF	16,2	3eev	False	30	2	Entrenada	-0.5	-0.4	91.67	100	94.12
56	FF	16,2	3eev	True	30	1	Entrenada	-0.4	-0.4	100	100	100
57	FF	16,2	3eve	False	30	2	Entrenada	-0.4	-0.4	81.82	83.33	82.35
58	FF	16,2	3eve	True	128	61	Detenida	-3.7	-0.6	0	33.33	11.76
59	FF	16,2	3vee	False	30	1	Entrenada	-0.5	-0.4	36.36	83.33	52.94
60	FF	16,2	3vee	True	30	6	Entrenada	-0.5	-0.4	45.45	83.33	58.82
61	FF	20,2	3eev	False	61	30	Entrenada	-1.3	-0.5	0	40	11.76
62	FF	20,2	3eev	True	30	8	Entrenada	-0.5	-0.4	83.33	80	82.35
63	FF	20,2	3eve	False	84	67	Detenida	-15.5	-0.4	0	33.33	11.76
64	FF	20,2	3eve	True	92	32	Entrenada	-0.9	-0.5	0	66.67	23.53
65	FF	20,2	3vee	False	30	1	Entrenada	-0.4	-0.4	90.91	100	94.12
66	FF	20,2	3vee	True	30	7	Entrenada	-0.6	-0.4	36.36	83.33	52.94
67	FF	4,4	3eev	False	61	22	Entrenada	-0.6	-0.5	16.67	40	23.53
68	FF	4,4	3eev	True	61	18	Entrenada	-0.5	-0.4	41.67	80	52.94
69	FF	4,4	3eve	False	30	3	Entrenada	-0.6	-0.4	36.36	83.33	52.94
70	FF	4,4	3eve	True	30	12	Entrenada	-0.6	-0.5	18.18	66.67	35.29
71	FF	4,4	3vee	False	61	29	Entrenada	-1	-0.4	0	66.67	23.53
72	FF	4,4	3vee	True	30	5	Entrenada	-0.5	-0.4	100	100	100
73	FF	8,4	3eev	False	30	2	Entrenada	-0.5	-0.4	83.33	100	88.24
74	FF	8,4	3eev	True	61	20	Entrenada	-0.9	-0.4	0	80	23.53
75	FF	8,4	3eve	False	30	11	Entrenada	-0.7	-0.5	18.18	50	29.41

Tabla 3.6: Resultados de los entrenamientos de las RNA del ejemplo de búsqueda (parte 4)

ID	Tipo	Neuronas	Método	Inicio	Iter.	Mejor iter.	Estado	Log. error ent.	Log. error val.	% error ent.	% Error val.	% error total
76	FF	8,4	3eve	True	92	37	Entrenada	-2.1	-0.5	0	33.33	11.76
77	FF	8,4	3vee	False	61	21	Entrenada	-1	-0.4	0	50	17.65
78	FF	8,4	3vee	True	30	5	Entrenada	-0.4	-0.4	100	100	100
79	FF	12,4	3eev	False	30	2	Entrenada	-0.5	-0.4	75	100	82.35
80	FF	12,4	3eev	True	30	12	Entrenada	-0.6	-0.4	33.33	100	52.94
81	FF	12,4	3eve	False	30	1	Entrenada	-0.5	-0.4	63.64	100	76.47
82	FF	12,4	3eve	True	123	48	Entrenada	-1.1	-0.5	0	50	17.65
83	FF	12,4	3vee	False	30	3	Entrenada	-0.5	-0.4	45.45	66.67	52.94
84	FF	12,4	3vee	True	30	4	Entrenada	-0.5	-0.4	63.64	100	76.47
85	FF	16,4	3eev	False	30	3	Entrenada	-0.5	-0.5	41.67	60	47.06
86	FF	16,4	3eev	True	100	62	Detenida	-6.1	-0.6	0	40	11.76
87	FF	16,4	3eve	False	30	1	Entrenada	-0.5	-0.4	90.91	100	94.12
88	FF	16,4	3eve	True	61	28	Entrenada	-0.8	-0.5	0	50	17.65
89	FF	16,4	3vee	False	30	7	Entrenada	-1	-0.4	0	50	17.65
90	FF	16,4	3vee	True	30	4	Entrenada	-0.5	-0.4	63.64	100	76.47
91	FF	20,4	3eev	False	30	1	Entrenada	-0.4	-0.4	91.67	100	94.12
92	FF	20,4	3eev	True	61	27	Entrenada	-0.9	-0.5	0	60	17.65
93	FF	20,4	3eve	False	61	17	Entrenada	-0.9	-0.6	0	33.33	11.76
94	FF	20,4	3eve	True	30	9	Entrenada	-0.7	-0.5	18.18	50	29.41
95	FF	20,4	3vee	False	30	5	Entrenada	-1.3	-0.4	0	50	17.65
96	FF	20,4	3vee	True	30	1	Entrenada	-0.4	-0.4	100	100	100
97	FF	4,6	3eev	False	490	332	Detenida	-1.8	-0.5	0	60	17.65
98	FF	4,6	3eev	True	30	12	Entrenada	-0.5	-0.4	66.67	100	76.47
99	FF	4,6	3eve	False	30	5	Entrenada	-0.5	-0.4	54.55	66.67	58.82
100	FF	4,6	3eve	True	61	20	Entrenada	-0.6	-0.4	18.18	83.33	41.18

Tabla 3.7: Resultados de los entrenamientos de las RNA del ejemplo de búsqueda (parte 5)

ID	Tipo	Neuronas	Método	Inicio	Iter.	Mejor iter.	Estado	Log. error ent.	Log. error val.	% error ent.	% error val.	% error total
101	FF	4,6	3vee	False	30	2	Entrenada	-0.5	-0.4	45.45	83.33	58.82
102	FF	4,6	3vee	True	30	3	Entrenada	-0.4	-0.4	100	100	100
103	FF	8,6	3eev	False	30	3	Entrenada	-0.6	-0.4	33.33	60	41.18
104	FF	8,6	3eev	True	86	31	Detenida	-0.9	-0.5	0	80	23.53
105	FF	8,6	3eve	False	123	60	Entrenada	-0.8	-0.4	0	100	35.29
106	FF	8,6	3eve	True	61	15	Entrenada	-0.8	-0.5	0	66.67	23.53
107	FF	8,6	3vee	False	30	3	Entrenada	-0.7	-0.3	0	100	35.29
108	FF	8,6	3vee	True	61	19	Entrenada	-0.7	-0.4	9.09	66.67	29.41
109	FF	12,6	3eev	False	30	13	Entrenada	-0.6	-0.4	16.67	80	35.29
110	FF	12,6	3eev	True	30	4	Entrenada	-0.4	-0.4	100	100	100
111	FF	12,6	3eve	False	30	3	Entrenada	-0.5	-0.4	63.64	100	76.47
112	FF	12,6	3eve	True	84	35	Detenida	-2.2	-0.6	0	16.67	5.88
113	FF	12,6	3vee	False	30	4	Entrenada	-0.8	-0.4	9.09	66.67	29.41
114	FF	12,6	3vee	True	61	19	Entrenada	-1.1	-0.4	0	66.67	23.53
115	FF	16,6	3eev	False	30	4	Entrenada	-0.6	-0.4	16.67	80	35.29
116	FF	16,6	3eev	True	30	3	Entrenada	-0.5	-0.4	91.67	100	94.12
117	FF	16,6	3eve	False	30	6	Entrenada	-0.6	-0.4	0	83.33	29.41
118	FF	16,6	3eve	True	61	27	Entrenada	-0.9	-0.5	0	50	17.65
119	FF	16,6	3vee	False	30	3	Entrenada	-0.7	-0.4	18.18	66.67	35.29
120	FF	16,6	3vee	True	30	4	Entrenada	-0.5	-0.4	72.73	100	82.35
121	FF	20,6	3eev	False	30	5	Entrenada	-0.6	-0.4	33.33	80	47.06
122	FF	20,6	3eev	True	61	19	Entrenada	-1	-0.5	0	40	11.76
123	FF	20,6	3eve	False	30	1	Entrenada	-0.5	-0.4	36.36	83.33	52.94
124	FF	20,6	3eve	True	79	64	Detenida	-15.7	-0.6	0	33.33	11.76
125	FF	20,6	3vee	False	30	2	Entrenada	-0.6	-0.4	9.09	83.33	35.29

Tabla 3.8: Resultados de los entrenamientos de las RNA del ejemplo de búsqueda (parte 6)

ID	Tipo	Neuronas	Método	Inicio	Iter.	Mejor iter.	Estado	Log. error ent.	Log. error val.	% error ent.	% error val.	% error total
126	FF	20,6	3vee	True	30	8	Entrenada	-0.6	-0.4	9.09	100	41.18
127	FF	4,8	3eev	False	61	26	Entrenada	-0.8	-0.3	0	80	23.53
128	FF	4,8	3eev	True	30	9	Entrenada	-0.5	-0.4	91.67	100	94.12
129	FF	4,8	3eve	False	30	14	Entrenada	-0.6	-0.3	36.36	83.33	52.94
130	FF	4,8	3eve	True	30	3	Entrenada	-0.4	-0.4	100	100	100
131	FF	4,8	3vee	False	30	6	Entrenada	-0.6	-0.4	9.09	66.67	29.41
132	FF	4,8	3vee	True	30	12	Entrenada	-0.5	-0.4	45.45	66.67	52.94
133	FF	8,8	3eev	False	216	93	Entrenada	-1.3	-0.4	0	60	17.65
134	FF	8,8	3eev	True	84	67	Detenida	-16.1	-0.5	0	20	5.88
135	FF	8,8	3eve	False	309	145	Entrenada	-1	-0.5	0	66.67	23.53
136	FF	8,8	3eve	True	61	24	Entrenada	-0.8	-0.5	0	66.67	23.53
137	FF	8,8	3vee	False	30	2	Entrenada	-0.6	-0.4	27.27	83.33	47.06
138	FF	8,8	3vee	True	30	5	Entrenada	-0.5	-0.4	45.45	100	64.71
139	FF	12,8	3eev	False	30	3	Entrenada	-0.6	-0.4	25	80	41.18
140	FF	12,8	3eev	True	61	30	Entrenada	-0.6	-0.4	25	80	41.18
141	FF	12,8	3eve	False	30	4	Entrenada	-0.7	-0.5	9.09	66.67	29.41
142	FF	12,8	3eve	True	30	5	Entrenada	-0.6	-0.4	45.45	100	64.71
143	FF	12,8	3vee	False	30	1	Entrenada	-0.5	-0.3	36.36	100	58.82
144	FF	12,8	3vee	True	30	11	Entrenada	-0.7	-0.4	9.09	66.67	29.41
145	FF	16,8	3eev	False	30	11	Entrenada	-1	-0.5	0	60	17.65
146	FF	16,8	3eev	True	61	29	Entrenada	-2.2	-0.5	0	60	17.65
147	FF	16,8	3eve	False	92	33	Entrenada	-2.5	-0.6	0	33.33	11.76
148	FF	16,8	3eve	True	123	61	Entrenada	-6.2	-0.5	0	50	17.65
149	FF	16,8	3vee	False	30	2	Entrenada	-0.7	-0.4	0	83.33	29.41
150	FF	16,8	3vee	True	30	2	Entrenada	-0.4	-0.4	90.91	100	94.12

Tabla 3.9: Resultados de los entrenamientos de las RNA del ejemplo de búsqueda (parte 7)

ID	Tipo	Neuronas	Método	Inicio	Iter.	Mejor iter.	Estado	Log. error ent.	Log. error val.	% error ent.	% error val.	% error total
151	FF	20,8	3cev	False	30	4	Entrenada	-0.7	-0.4	16.67	100	41.18
152	FF	20,8	3cev	True	129	78	Detenida	-15	-0.5	0	40	11.76
153	FF	20,8	3cve	False	30	4	Entrenada	-0.6	-0.4	9.09	100	41.18
154	FF	20,8	3cve	True	30	7	Entrenada	-0.6	-0.4	27.27	100	52.94
155	FF	20,8	3vce	False	30	3	Entrenada	-1.1	-0.4	0	83.33	29.41
156	FF	20,8	3vce	True	30	4	Entrenada	-0.5	-0.4	81.82	100	88.24
157	FF	4,10	3cev	False	30	2	Entrenada	-0.6	-0.4	33.33	80	47.06
158	FF	4,10	3cev	True	102	72	Detenida	-18	-0.5	0	40	11.76
159	FF	4,10	3cve	False	559	327	Detenida	-2.5	-0.4	0	66.67	23.53
160	FF	4,10	3cve	True	61	28	Entrenada	-0.9	-0.7	9.09	16.67	11.76
161	FF	4,10	3vce	False	123	59	Entrenada	-1	-0.4	0	66.67	23.53
162	FF	4,10	3vce	True	30	1	Entrenada	-0.4	-0.4	100	100	100
163	FF	8,10	3cev	False	30	3	Entrenada	-0.6	-0.4	25	60	35.29
164	FF	8,10	3cev	True	104	64	Detenida	-4.9	-0.7	0	20	5.88
165	FF	8,10	3cve	False	30	14	Entrenada	-0.9	-0.4	0	66.67	23.53
166	FF	8,10	3cve	True	61	27	Entrenada	-1.1	-0.5	0	33.33	11.76
167	FF	8,10	3vce	False	30	2	Entrenada	-0.7	-0.4	9.09	83.33	35.29
168	FF	8,10	3vce	True	30	8	Entrenada	-0.5	-0.4	45.45	83.33	58.82
169	FF	12,10	3cev	False	253	135	Detenida	-4.2	-0.8	0	60	17.65
170	FF	12,10	3cev	True	61	22	Entrenada	-1.1	-0.4	0	60	17.65
171	FF	12,10	3cve	False	92	34	Entrenada	-0.8	-0.3	9.09	83.33	35.29
172	FF	12,10	3cve	True	30	2	Entrenada	-0.4	-0.4	100	100	100
173	FF	12,10	3vce	False	92	40	Entrenada	-1.4	-0.3	0	100	35.29
174	FF	12,10	3vce	True	30	1	Entrenada	-0.4	-0.4	100	100	100
175	FF	16,10	3cev	False	30	3	Entrenada	-0.6	-0.4	33.33	80	47.06

Tabla 3.10: Resultados de los entrenamientos de las RNA del ejemplo de búsqueda (parte 8)

ID	Tipo	Neuronas	Método	Inicio	Iter.	Mejor iter.	Estado	Log. error ent.	Log. error val.	% error ent.	% error val.	% error total
176	FF	16,10	3cev	True	61	24	Entrenada	-1	-0.5	0	40	11.76
177	FF	16,10	3cve	False	30	3	Entrenada	-0.8	-0.3	9.09	100	41.18
178	FF	16,10	3cve	True	61	27	Entrenada	-0.9	-0.5	0	33.33	11.76
179	FF	16,10	3vce	False	30	5	Entrenada	-0.8	-0.4	0	66.67	23.53
180	FF	16,10	3vce	True	30	3	Entrenada	-0.5	-0.4	100	100	100
181	FF	20,10	3cev	False	30	6	Entrenada	-0.8	-0.3	0	80	23.53
182	FF	20,10	3cev	True	104	83	Detenida	-15.5	-0.5	0	40	11.76
183	FF	20,10	3cve	False	154	75	Entrenada	-2.5	-0.8	0	16.67	5.88
184	FF	20,10	3cve	True	61	24	Entrenada	-2	-0.6	0	33.33	11.76
185	FF	20,10	3vce	False	30	6	Entrenada	-1.3	-0.3	0	66.67	23.53
186	FF	20,10	3vce	True	30	9	Entrenada	-0.8	-0.4	0	66.67	23.53

Tabla 3.11: Mejores RNA del ejemplo de búsqueda, ordenadas por el logaritmo del error de validación

ID	Tipo	Neuronas	Método	Inicio	Iter.	Mejor iter.	Estado	Log. error ent.	Log. error val.	% error ent.	% error val.	% error total
7	FF	4,0	3cev	False	61	26	Entrenada	-1	-0.8	0	0	0
169	FF	12,10	3cev	False	253	135	Detenida	-4.2	-0.8	0	60	17.65
183	FF	20,10	3eve	False	154	75	Entrenada	-2.5	-0.8	0	16.67	5.88
16	FF	8,0	3eve	True	79	41	Detenida	-6.7	-0.7	0	16.67	5.88
22	FF	12,0	3eve	True	97	92	Detenida	-15.7	-0.7	0	33.33	11.76
28	FF	16,0	3eve	True	89	38	Detenida	-2.1	-0.7	0	16.67	5.88
31	FF	20,0	3cev	False	30	11	Entrenada	-1.4	-0.7	0	20	5.88
44	FF	8,2	3cev	True	61	30	Entrenada	-0.7	-0.7	25	20	23.53
160	FF	4,10	3eve	True	61	28	Entrenada	-0.9	-0.7	9.09	16.67	11.76
164	FF	8,10	3cev	True	104	64	Detenida	-4.9	-0.7	0	20	5.88

Tabla 3.12: Mejores RNA del ejemplo de búsqueda, ordenadas según el porcentaje de error de validación

ID	Tipo	Neuronas	Método	Inicio	Iter.	Mejor iter.	Estado	Log. error ent.	Log. error val.	% error ent.	% error val.	% error total
7	**FF**	**4,0**	**3eev**	**False**	**61**	**26**	**Entrenada**	**-1**	**-0.8**	**0**	**0**	**0**
16	FF	8,0	3eve	True	79	41	Detenida	-6.7	-0.7	0	16.67	5.88
28	FF	16,0	3eve	True	89	38	Detenida	-2.1	-0.7	0	16.67	5.88
112	FF	12,6	3eve	True	84	35	Detenida	-2.2	-0.6	0	16.67	5.88
160	FF	4,10	3eve	True	61	28	Entrenada	-0.9	-0.7	9.09	16.67	11.76
183	FF	20,10	3eve	False	154	75	Entrenada	-2.5	-0.8	0	16.67	5.88
8	FF	4,0	3eev	True	92	33	Entrenada	-0.7	-0.5	16.67	20	17.65
31	FF	20,0	3eev	False	30	11	Entrenada	-1.4	-0.7	0	20	5.88
38	FF	4,2	3eev	True	91	41	Entrenada	-2.5	-0.6	0	20	5.88
44	FF	8,2	3eev	True	61	30	Entrenada	-0.7	-0.7	25	20	23.53

en la iteración 26ª. Téngase en cuenta que se ha conseguido entrenar dicha RNA con 17 casos y 80 variables, en 26 iteraciones, pero también debería llamar la atención que esta RNA presenta tanto un porcentaje de error de entrenamiento como un porcentaje de error de validación del 0 %. Esto podría explicarse si la RNA hubiera modelizado perfectamente la ley que subyace bajo el fenómeno que ha producido los datos. Aunque no es fácil demostrar la afirmación anterior sin conocer el origen de los datos, esta posibilidad es muy interesante, pues se habría conseguido llegar a clasificar correctamente (también desde el punto de vista determinístico o exacto) un conjunto de datos en 5 categorías mediante una RNA relativamente sencilla. Además de las características anteriormente mencionadas, dicha RNA, en concreto, viene dada por el vector X de los datos conocidos y por la matriz pesos W definida como:

$$W = \begin{pmatrix}
2{,}10367 & 1{,}03435 & -0{,}53167 & -2{,}49828 \\
-1{,}40292 & 0{,}453688 & -0{,}206948 & -0{,}965436 \\
-0{,}713408 & -0{,}378824 & -0{,}895444 & 1{,}1408 \\
-0{,}129773 & 0{,}0080916 & -0{,}426379 & -2{,}68489 \\
0{,}0809884 & 0{,}0480173 & -0{,}644341 & -0{,}9603 \\
0{,}649009 & 0{,}292174 & 0{,}808398 & -1{,}5959 \\
-0{,}221986 & 0{,}24802 & 0{,}456513 & -2{,}08658 \\
0{,}330777 & 0{,}0727023 & 0{,}575113 & 0{,}345374 \\
-0{,}367415 & 0{,}259473 & 0{,}167451 & -2{,}17027 \\
-0{,}231296 & 0{,}649006 & 0{,}538398 & -0{,}665618 \\
-1{,}88855 & -0{,}236871 & 1{,}43081 & -1{,}06547 \\
0{,}14777 & -1{,}09805 & 0{,}36091 & -1{,}21588 \\
-0{,}316453 & 0{,}027505 & -0{,}0588953 & 0{,}864805 \\
-0{,}818052 & 0{,}308828 & -0{,}948352 & -0{,}551917 \\
-1{,}11236 & 0{,}00405728 & -0{,}266778 & -1{,}36273 \\
-1{,}16195 & -0{,}0645851 & -0{,}135068 & 2{,}42171 \\
0{,}985112 & -0{,}0999166 & -0{,}680898 & 0{,}348493 \\
-1{,}22906 & 0{,}653133 & -0{,}551653 & -1{,}03922 \\
-0{,}321671 & -0{,}00523073 & 0{,}567284 & -0{,}299924 \\
-0{,}678208 & -0{,}0379668 & 2{,}13876 & -0{,}922988 \\
1{,}06961 & -1{,}09197 & 0{,}872837 & 2{,}07701 \\
1{,}07337 & -0{,}477786 & 0{,}465172 & 0{,}146879 \\
-0{,}796171 & 0{,}00847426 & 1{,}02953 & 1{,}47409 \\
1{,}98752 & 0{,}340576 & -0{,}313572 & -1{,}26857 \\
-1{,}15034 & 0{,}117283 & -0{,}233019 & 1{,}36395 \\
0{,}3137 & -0{,}905426 & -1{,}14586 & -2{,}01982 \\
-0{,}97664 & -0{,}58026 & 0{,}40495 & 0{,}687624 \\
0{,}689142 & -1{,}02507 & -1{,}25217 & 1{,}12712 \\
-0{,}446235 & 0{,}972726 & -0{,}555515 & -1{,}9742 \\
-0{,}831559 & 0{,}0629595 & 0{,}155396 & -0{,}603315 \\
-0{,}190477 & -0{,}767902 & 1{,}64117 & 3{,}05972 \\
-0{,}00351516 & 0{,}560536 & -0{,}688172 & 0{,}220992 \\
0{,}0608371 & 0{,}26705 & -0{,}0316905 & -0{,}44706 \\
-0{,}848247 & 0{,}0122545 & 0{,}665829 & 1{,}74597 \\
-0{,}713534 & 0{,}741944 & -0{,}17071 & 0{,}682114 \\
-0{,}301807 & 0{,}316228 & -0{,}0234985 & 0{,}783304 \\
0{,}607516 & 1{,}04609 & 0{,}0312204 & 0{,}948337 \\
-0{,}277629 & -0{,}653727 & -0{,}64156 & -1{,}78255 \\
-0{,}364903 & -0{,}0835399 & 0{,}234206 & -1{,}44409 \\
0{,}0302884 & 0{,}344384 & 0{,}381009 & -0{,}179892 \\
-0{,}722453 & -0{,}0698637 & -0{,}771225 & 1{,}62652 \\
-0{,}290582 & -1{,}43862 & 1{,}22915 & 0{,}129091 \\
-0{,}984511 & -0{,}372289 & -1{,}0274 & -1{,}39716 \\
1{,}52446 & 0{,}0263177 & 1{,}11386 & -0{,}130955 \\
1{,}34409 & 0{,}4052 & 0{,}241248 & -2{,}24668 \\
-0{,}817441 & -0{,}921823 & -0{,}632187 & 1{,}53858 \\
0{,}938988 & -0{,}659819 & 0{,}607736 & 1{,}52819 \\
0{,}219011 & -0{,}454196 & 0{,}347103 & 0{,}519259 \\
-0{,}457287 & 0{,}942166 & -0{,}161855 & -1{,}54484 \\
0{,}34851 & 0{,}574008 & -1{,}40319 & 0{,}993575 \\
-0{,}814195 & 4{,}81079 & -3{,}83278 & 3{,}68764
\end{pmatrix}$$

Como ya se ha comentado, la estructura topológica subyacente es la $G = (V, E)$ de la Figura 3.5, la función neuronal asociada a la estructura topológica anterior viene dada por la función sigmoide y el sistema de aprendizaje ha utilizado el método del gradiente conjugado.

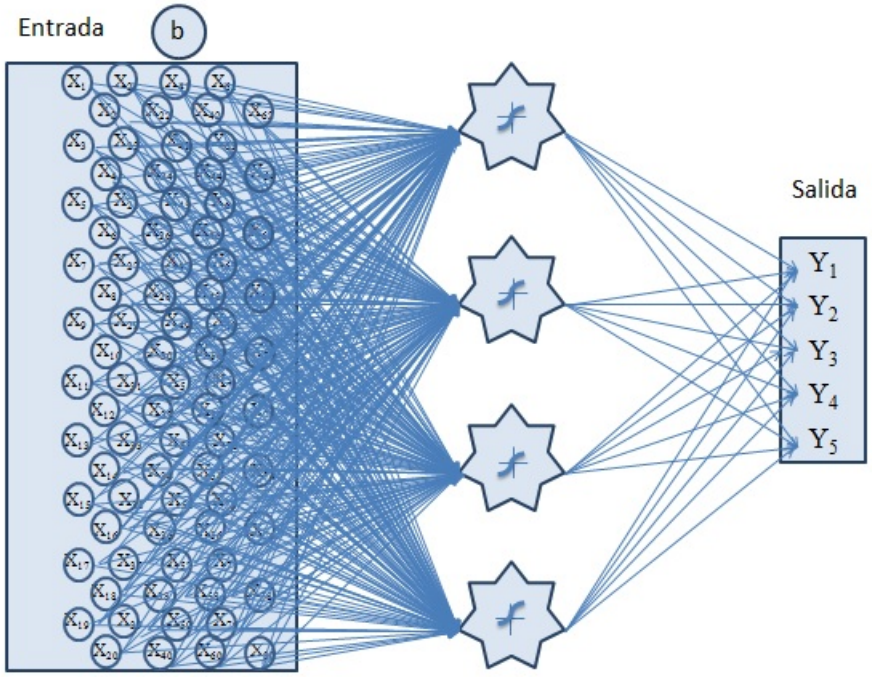


Figura 3.5: Ejemplo de RNA adecuada, seleccionada por el programa informático

Al contrario de lo que ocurre con la RNA representada en la Figura 3.5, hay otras redes que también han sido entrenadas durante el proceso y que cometen un error de entrenamiento superior al 90 % (como, por ejemplo, la 91ª). Este suceso puede servirnos para resumir este apartado y señalar la importancia de establecer correctamente la estructura topológica subyacente de una RNA (y el resto de características de la misma) a la hora de afrontar un problema, para elegir

una RNA que realmente ajuste el comportamiento de los datos o que modelice el fenómeno que los genera.

3.3. Reducción de parámetros y tiempo computacional

Como ya se ha comentado con anterioridad, uno de los grandes inconvenientes de la utilización de las RNA es el elevado número de parámetros que se tienen que calcular, ya que este valor viene influido por el número de variables analizadas, por la naturaleza de los casos del estudio, por la estructura topológica subyacente elegida y hasta por el valor inicial de las conexiones existentes.

Un objetivo parcial inicial de este trabajo consistía en buscar un procedimiento automático para decidir si una determinada variable debe o no entrar a formar parte de un modelo. En principio, una RNA elimina una variable de un modelo asignándole ceros al o a los pesos correspondientes a dicha variable. Sin embargo, no todos los algoritmos de entrenamiento garantizan la aparición de ceros en los lugares que serían más convenientes. La eliminación de una variable puede, desde este punto de vista, ser considerada un tipo de reducción de parámetros, aunque podemos llegar a alcanzar una reducción incluso más interesante.

Esta sección la dedicamos a presentar una pequeña propuesta de mejora, referida a la estructura topológica, concretamente utilizando los pesos iniciales condicionados a las relaciones intrínsecas entre las variables predictoras; es decir, utilizando la correlación fuerte que existe en la mayoría de los casos entre las variables independientes (y también con la variable dependiente). Este hecho, además de ser muy usual en la práctica, es relevante para llevar a cabo cualquier análisis estadístico.

La propuesta de esta sección tiene un especial sentido cuando se hallan subconjuntos dentro del conjunto de variables estudiadas; es decir, siempre que exista una relación entre variables y ellas se puedan agrupar en subconjuntos de características medidas por diferentes variables en el mismo análisis. Recíprocamente, también se puede considerar cuando se constata que hay dos tipos de relaciones entre las variables, que producen grupos más o menos homogéneos de variables

que, entre ellos, son relativamente independientes (con una reducida relación inter-grupos).

En resumen, la idea de fondo se basa en que, cuando se encuentren relaciones entre las variables o conjuntos de características de una misma índole, consideramos que sería interesante que la relación entre estas características se estudiara primero por separado (análisis de las relaciones intra-grupos de variables) para realizar un análisis posterior con el peso correspondiente a cada una de las variables. Esta estrategia está bastante relacionada con el modo de proceder más natural cuando se trata, por ejemplo, de definir indicadores multidimensionales: por una parte, se analizan características más o menos homogéneas; después se incorporan las diferentes "características homogeneizadas" en un único valor que comprenda toda la información. Es decir, si tenemos variables que estudian la característica 1 y otras variables que estudian la característica 2, proponemos realizar un análisis previo de las características por separado antes de reunir todas las variables. Este hecho se puede llevar a cabo simplemente con la definición de algunos pesos iniciales nulos en una RNA; esto es, en la matriz de pesos iniciales, se asignará un 0 a las conexiones entre las variables de entrada y los subgrupos (denotados por nodos en la primera capa oculta) que correspondan a características distintas a las de la propia variable. A continuación se introduce un ejemplo para tratar de explicar más claramente esta propuesta metodológica.

3.3.1. Ejemplo de simplificación de la estructura topológica

En el siguiente ejemplo (de clasificación de un conjunto de individuos en 10 grupos distintos) se puede observar que con una simplificación en los valores de los pesos iniciales de la estructura topológica se puede reducir tanto el tiempo computacional como el error que se comete. El conjunto que hemos utilizado en este ejemplo está compuesto por 493 individuos, de los cuales conocemos 87 varia-

bles de carácter binario; estas variables realmente se refieren a 20 características de cada individuo. Luego, atendiendo a que existe una relación entre las distintas 87 variables, ya que existen conjuntos de variables que observan la misma característica, se propone el siguiente análisis, que tiene en cuenta las variables que están relacionadas entre sí y las que no (es decir, que pueden considerarse independientes).

En primer lugar, realizamos un primer análisis con el conjunto de variables en general como vector de entrada y, como vector de salida, el grupo donde se caracteriza el individuo, atendiendo a que existen 10 grupos distintos.

Para realizar la primera clasificación (sin simplificar la RNA), utilizamos una RNA con dos capas ocultas, de modo que en la primera capa oculta utilizamos 20 neuronas y en la segunda capa existen 4 neuronas. Es decir, primero utilizamos una matriz W de pesos iniciales aleatoria. Una vez determinada esta matriz aleatoria de pesos iniciales, realizamos el entrenamiento de la RNA, obteniendo un error de $0,21303$ y con un tiempo computacional de $2287,22$ segundos.

En la Figura 3.6 se representa la RNA utilizada, con la multiplicidad de pesos iniciales. En el dibujo no se ha podido incluir el gráfico completo, por existir un número muy elevado de conexiones iniciales, ya que por cada conjunto de variables existen 20 conexiones, todas ellas distintas de cero, porque por definición los valores de la matriz inicial de pesos aleatoria son todos distintos de cero.

Una vez realizado este primer análisis, decidimos realizar una mejora que supone un cambio en la matriz de pesos iniciales. Utilizando la anterior matriz de pesos aleatoria, se actualizaron sus elementos, introduciendo un 0 en los valores correspondientes a conexiones entre distintos conjuntos de variables. Para ello, utilizamos 20 neuronas en la primera capa, coincidiendo con las características iniciales; es decir, las 87 variables se convierten en 20 características. Luego, utilizamos los valores de los pesos como 0 en todas las variables que no estaban

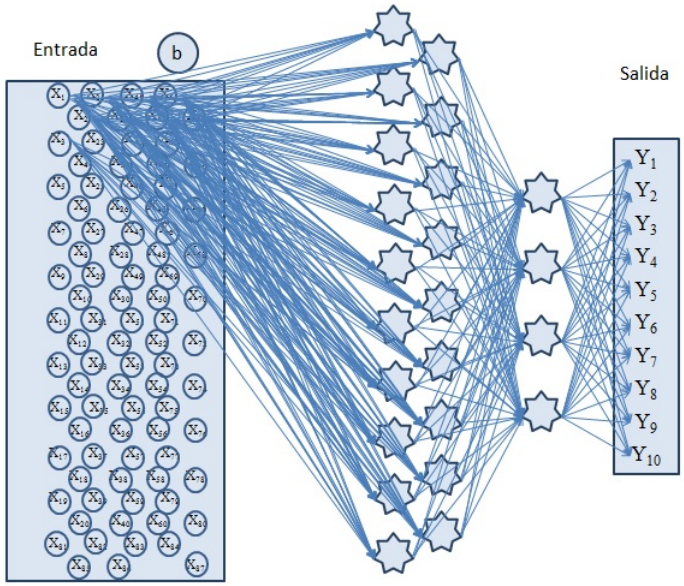


Figura 3.6: Ejemplo de la RNA utilizada, con el conjunto completo de conexiones iniciales

relacionadas con la característica estudiada; con ello, se reduce el número de conexiones en la primera capa, como se puede ver en la Figura 3.7.

A continuación describimos la matriz de pesos W (por cajas):

$$W = \left(\begin{array}{c|c|c|c} A_{11} & \Theta & \Theta & \Theta \\ \hline \Theta & A_{22} & \Theta & \Theta \\ \hline \Theta & \Theta & A_{33} & \Theta \\ \hline \Theta & \Theta & \Theta & A_{44} \\ \hline B_1 & B_2 & B_3 & B_4 \end{array}\right)$$

Las submatrices componentes vienen definidas de la siguiente forma (se utiliza la notación habitual de las matrices por cajas para permitir su presentación en un espacio razonable y el punto decimal anglosajón por resultar ser las salidas de computación de Mathematica):

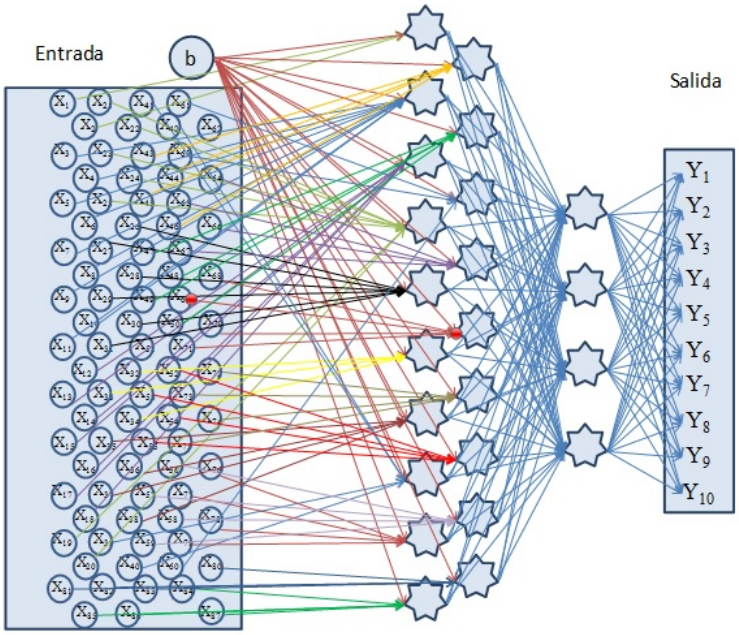


Figura 3.7: Ejemplo de la RNA utilizada tras la simplificación

$$B_1 = \begin{pmatrix} 3{,}16286 & -8{,}2659 \end{pmatrix}$$

$$B_2 = \begin{pmatrix} -5{,}06851 & -4{,}43856 \end{pmatrix}$$

$$B_3 = \begin{pmatrix} 3{,}78219 & 5{,}67951 & 3{,}8063 & 4{,}00916 & 7{,}47444 & -1{,}50918 \end{pmatrix}$$

$$B_4 = \left(\begin{array}{c|c} B_{41} & B_{42} \end{array} \right)$$

$$B_{41} = \begin{pmatrix} -4{,}89056 & 4{,}26967 & 8{,}31506 & -6{,}04258 & 1{,}04234 \end{pmatrix}$$

$$B_{42} = \begin{pmatrix} 1{,}17592 & -5{,}26554 & 1{,}60746 & 0{,}316853 & -7{,}26502 \end{pmatrix}$$

$$A_{11} = \begin{pmatrix} -0{,}167612 & -0{,}588207 \\ 0{,}359643 & 1{,}00778 \\ 0{,}97252 & 0{,}0723727 \\ 1{,}42089 & 1{,}04107 \\ -1{,}55405 & -1{,}09133 \\ 0{,}475649 & -0{,}103148 \\ -0{,}328503 & -0{,}459862 \\ 0{,}0326097 & -0{,}597328 \\ -0{,}0467303 & 0{,}781025 \\ -0{,}918743 & 0{,}72482 \\ -0{,}560917 & 1{,}20603 \end{pmatrix} \quad A_{22} = \begin{pmatrix} -0{,}843663 & 1{,}49478 \\ -0{,}919201 & 0{,}576119 \\ 0{,}358657 & 1{,}10645 \\ 1{,}34079 & -0{,}844404 \\ 0{,}255927 & 0{,}664771 \\ -1{,}70392 & -1{,}29931 \\ -0{,}978179 & -0{,}147957 \\ 0{,}85119 & -0{,}876774 \\ 0{,}78562 & -1{,}43592 \\ -0{,}656038 & -1{,}54245 \\ 1{,}03439 & 1{,}15826 \\ -0{,}593106 & -1{,}9091 \\ 1{,}34766 & 0{,}00133069 \\ -1{,}50564 & -0{,}748944 \end{pmatrix}$$

$$A_{33} = \begin{pmatrix} -1{,}5361 & -0{,}849206 & 0{,}0544839 & 0{,}938293 & 1{,}95776 & 0{,}992775 \\ -0{,}991492 & 1{,}08242 & -1{,}17977 & 0{,}90801 & -0{,}725366 & -0{,}97258 \\ 1{,}02327 & 0{,}267673 & -0{,}795519 & -0{,}912363 & 1{,}22057 & 0{,}420324 \\ -0{,}280918 & 1{,}45876 & 0{,}980631 & -0{,}9686 & 0{,}285978 & 2{,}0399 \\ 1{,}59991 & -0{,}717666 & -0{,}628095 & 0{,}5762 & 0{,}0231366 & 0{,}626053 \\ 0{,}0231086 & 1{,}45607 & 1{,}26544 & -0{,}896608 & 0{,}598987 & -1{,}209 \\ -0{,}964154 & -0{,}485116 & -0{,}501845 & -0{,}200887 & 0{,}622178 & -0{,}942923 \\ -0{,}717755 & 0{,}0458273 & 1{,}08431 & -0{,}941104 & 0{,}933991 & 0{,}84902 \\ 1{,}92053 & -1{,}73009 & -0{,}0542549 & -0{,}458694 & 0{,}966589 & -1{,}96454 \\ 1{,}77268 & 1{,}01287 & -0{,}486216 & -0{,}446981 & -0{,}247847 & -0{,}884507 \\ 1{,}53546 & 0{,}872559 & 0{,}134304 & 0{,}933291 & -0{,}939053 & -1{,}21721 \\ 0{,}0547852 & -0{,}035718 & -1{,}37902 & 0{,}645373 & -0{,}121654 & 0{,}365699 \\ 0{,}760745 & -0{,}321741 & 1{,}41841 & 0{,}737241 & -0{,}759337 & 0{,}777698 \\ 1{,}89352 & -0{,}347933 & -0{,}912302 & 0{,}152823 & -0{,}393569 & 0{,}92209 \\ -1{,}04097 & 0{,}170883 & 0{,}775508 & -0{,}437466 & -1{,}34145 & 0{,}98582 \\ -0{,}507401 & 0{,}271587 & -0{,}155857 & 0{,}0456536 & 1{,}61658 & -1{,}08439 \\ 0{,}227499 & 0{,}13639 & -0{,}674046 & 0{,}162256 & -0{,}181102 & -0{,}53454 \\ 1{,}06991 & -1{,}63997 & -0{,}890523 & 0{,}74706 & 0{,}629247 & -0{,}742083 \\ 0{,}746956 & 0{,}531545 & -0{,}97821 & 0{,}285377 & 1{,}13374 & 1{,}37185 \\ 1{,}66391 & -0{,}463031 & 1{,}10552 & -0{,}740085 & -0{,}32172 & 0{,}748318 \\ -1{,}9174 & 0{,}753485 & -0{,}782198 & -1{,}4208 & -1{,}20412 & -1{,}218 \\ -0{,}719091 & -0{,}125384 & 1{,}05945 & -0{,}904525 & -0{,}487744 & -0{,}702805 \\ -1{,}53569 & -0{,}164552 & -2{,}05638 & -1{,}66458 & 1{,}12159 & 0{,}921055 \\ 0{,}88308 & 1{,}73702 & 0{,}283502 & -1{,}24458 & 0{,}758833 & 0{,}296858 \\ 1{,}47582 & 1{,}49333 & 0{,}727175 & -0{,}290595 & -0{,}0677145 & 1{,}56719 \end{pmatrix}$$

$$A_{44} = \left(\begin{array}{c|c} A_{441} & A_{442} \end{array} \right)$$

$$A_{441} = \begin{pmatrix}
0{,}254307 & 1{,}16969 & -0{,}201636 & 0{,}141031 & 0{,}555143 \\
-0{,}91033 & -0{,}342234 & -0{,}843971 & -1{,}23351 & -0{,}846973 \\
-0{,}159439 & 1{,}16465 & 0{,}138588 & 0{,}988973 & 0{,}523789 \\
-1{,}04277 & -0{,}372793 & -0{,}585755 & 0{,}177439 & 0{,}893483 \\
0{,}733508 & -0{,}318716 & 1{,}0147 & 1{,}72128 & 0{,}338905 \\
-0{,}3479 & -0{,}0858443 & 0{,}0948916 & 0{,}194762 & 0{,}981894 \\
-0{,}990395 & -1{,}20624 & 1{,}64439 & 0{,}0455011 & -0{,}807233 \\
-0{,}416124 & 1{,}39271 & -0{,}559287 & 0{,}61026 & 0{,}647158 \\
-0{,}644441 & -1{,}35872 & 0{,}992811 & -0{,}755063 & 0{,}9046 \\
-0{,}260118 & -0{,}675494 & 0{,}388573 & -0{,}676386 & 1{,}16123 \\
-0{,}0640807 & -0{,}613743 & -0{,}421934 & 0{,}607517 & 0{,}388401 \\
-0{,}143414 & -0{,}763679 & 0{,}48326 & 0{,}678727 & -0{,}482697 \\
-0{,}0564995 & 1{,}10385 & -0{,}145072 & -1{,}17669 & 0{,}796049 \\
-0{,}773042 & -0{,}268621 & -1{,}08419 & 1{,}68758 & -1{,}32911 \\
0{,}555796 & 2{,}09323 & 0{,}665418 & 1{,}10863 & 0{,}814885 \\
-0{,}396451 & -0{,}151793 & -0{,}733441 & 0{,}341192 & 0{,}486777 \\
1{,}36782 & -0{,}784176 & -1{,}01367 & 0{,}307767 & 1{,}2582 \\
1{,}38328 & -0{,}58544 & -0{,}673247 & -0{,}593492 & -0{,}631656 \\
0{,}0409424 & -0{,}865929 & 1{,}5512 & -1{,}2051 & -0{,}87 \\
-0{,}649082 & -1{,}29408 & 0{,}250926 & 1{,}26254 & 0{,}497898 \\
-0{,}315553 & 0{,}0706356 & 0{,}388013 & -0{,}446805 & 0{,}803128 \\
1{,}32492 & 0{,}287566 & 0{,}568639 & 0{,}446734 & -1{,}67785 \\
0{,}0975708 & -1{,}93102 & -0{,}878265 & 0{,}917561 & -1{,}38482 \\
-1{,}36268 & -0{,}30683 & -0{,}21029 & 2{,}34965 & -1{,}03354 \\
-0{,}466606 & -1{,}26204 & 0{,}72967 & 1{,}10437 & 0{,}011101 \\
-1{,}0644 & -0{,}690387 & 0{,}0569565 & -0{,}590099 & 0{,}461643 \\
-0{,}831026 & -0{,}916291 & -1{,}15076 & 0{,}648808 & 0{,}452513 \\
-0{,}835393 & -0{,}871863 & 0{,}408727 & 1{,}35652 & 0{,}89138 \\
0{,}969019 & 0{,}481211 & -0{,}683246 & -0{,}443538 & 0{,}895273 \\
-1{,}21257 & 1{,}32456 & 1{,}02038 & 0{,}725077 & 0{,}0133198 \\
-1{,}27063 & 1{,}48601 & -0{,}573891 & 1{,}30944 & -1{,}02042 \\
-0{,}334454 & 1{,}09806 & -0{,}253992 & -0{,}159006 & -0{,}0274717 \\
-1{,}20721 & 0{,}736713 & -0{,}454821 & -0{,}360074 & 0{,}861661 \\
-0{,}851656 & -1{,}51566 & -0{,}333728 & 0{,}412784 & -0{,}810562 \\
-0{,}316998 & 1{,}04149 & 0{,}82679 & 1{,}38985 & -0{,}033785 \\
1{,}09326 & -0{,}860149 & 0{,}566201 & 1{,}00395 & 0{,}0181505 \\
-0{,}682199 & 1{,}21172 & -1{,}05389 & 0{,}122071 & 1{,}31368
\end{pmatrix}$$

$$A_{442} = \begin{pmatrix}
0{,}785551 & 0{,}58182 & 1{,}09751 & 0{,}156965 & -0{,}512943 \\
-1{,}18121 & -0{,}84306 & -0{,}770842 & -0{,}345717 & -1{,}12091 \\
0{,}081721 & -0{,}156392 & 0{,}891624 & -0{,}210655 & 1{,}35836 \\
1{,}63007 & 0{,}321329 & 0{,}458976 & 1{,}31743 & -0{,}514864 \\
-1{,}02407 & -1{,}15874 & -0{,}0920051 & 0{,}496773 & 0{,}445855 \\
-0{,}990481 & 0{,}679591 & 0{,}00979464 & 0{,}25149 & -0{,}230705 \\
0{,}512151 & 1{,}8537 & 0{,}703725 & -0{,}563514 & -2{,}14299 \\
0{,}292986 & 0{,}0656556 & -0{,}757684 & -0{,}475372 & 0{,}413707 \\
0{,}46169 & 1{,}89897 & 0{,}714979 & 0{,}109151 & 0{,}232815 \\
1{,}56387 & -1{,}22774 & 1{,}26675 & -0{,}566432 & -0{,}665393 \\
0{,}604548 & -0{,}103021 & -0{,}782621 & -0{,}369388 & -0{,}565955 \\
-1{,}02674 & -0{,}826647 & 0{,}0357452 & 1{,}20299 & 0{,}861923 \\
0{,}292887 & -0{,}723674 & -0{,}646691 & -0{,}386679 & -1{,}71001 \\
-0{,}81841 & -0{,}366137 & -0{,}814495 & -0{,}724751 & -0{,}124109 \\
-0{,}499069 & -0{,}99842 & -0{,}180707 & -0{,}568694 & -0{,}435345 \\
0{,}740695 & -0{,}47407 & -0{,}970015 & 0{,}445341 & 0{,}0514188 \\
-1{,}10059 & 0{,}6917 & -0{,}347868 & 0{,}844133 & -0{,}965422 \\
0{,}658452 & -0{,}0735071 & -0{,}407312 & -0{,}887145 & 1{,}52761 \\
-1{,}42787 & 1{,}09125 & 0{,}330306 & 1{,}85414 & -0{,}30129 \\
-1{,}56141 & 0{,}754821 & -0{,}0900816 & -0{,}954795 & -1{,}00369 \\
-1{,}21049 & 0{,}19716 & 0{,}987087 & 0{,}574799 & -1{,}48965 \\
-0{,}665643 & 0{,}209694 & -1{,}06862 & 0{,}0743113 & -1{,}55157 \\
-0{,}760792 & -0{,}731809 & -0{,}927037 & 1{,}01093 & 1{,}41358 \\
0{,}0767931 & 1{,}21189 & -0{,}945516 & 1{,}07988 & -0{,}647364 \\
2{,}31936 & 1{,}69503 & 0{,}786906 & -1{,}13651 & -0{,}486689 \\
-0{,}273315 & -1{,}40187 & -0{,}223234 & -0{,}454431 & -0{,}0284079 \\
1{,}72128 & 0{,}86967 & -0{,}464217 & 0{,}206073 & 0{,}750818 \\
-0{,}782476 & -0{,}419436 & -0{,}539511 & 0{,}261083 & -1{,}23145 \\
-0{,}485256 & -0{,}0940754 & 0{,}652801 & 0{,}870889 & -1{,}34358 \\
-1{,}1996 & 0{,}180332 & -0{,}855705 & -1{,}40377 & 1{,}40822 \\
0{,}974876 & 0{,}15735 & 0{,}946578 & 1{,}09747 & 1{,}2886 \\
0{,}096732 & 1{,}61521 & -0{,}37052 & 0{,}797678 & 1{,}77013 \\
-1{,}27982 & 1{,}40286 & -0{,}827135 & -0{,}0777558 & -1{,}21954 \\
-0{,}135897 & -0{,}489901 & 1{,}48955 & -0{,}34617 & -0{,}320212 \\
-0{,}930327 & 0{,}629762 & 0{,}825809 & 0{,}736643 & 1{,}46843 \\
0{,}309311 & 1{,}27426 & -1{,}32671 & -1{,}43944 & 1{,}26955 \\
-0{,}238233 & 1{,}04748 & 0{,}401848 & -0{,}185809 & -1{,}16374
\end{pmatrix}$$

La matriz W por cajas anterior la utilizamos como matriz inicial de pesos para el entrenamiento de la RNA. Por lo demás, este análisis consiste en el mismo entrenamiento que se realizó anteriormente, obteniéndose los siguientes resultados: un error de $0,2030$ y un tiempo computacional de $2077,95$ segundos. Como se puede apreciar, se produce una reducción tanto en el error computacional como en el tiempo que se tarda en calcular los parámetros de la RNA entrenada. En concreto, el error se reduce de un 21,5 % a un 20,3 % y el tiempo pasa de $2287,22$ a $2077,95$ (lo que serían 210 segundos menos; es decir, 3 minutos y medio de ahorro).

3.4. Propuesta de delimitación de técnicas

Al utilizar RNA en diferentes situaciones (de tratamiento de datos reales), un problema que hemos detectado es que la distribución que siguen los datos determina qué tipo de RNA es más adecuada para analizar el problema. Sin embargo, las características de la distribución rara vez se analizan automáticamente y un investigador que no tenga mucha experiencia en RNA difícilmente sabrá encontrar la RNA más adecuada a las características de sus datos. De hecho, es muy frecuente encontrar trabajos (incluso publicados en revistas prestigiosas) en los que se utilizan técnicas no apropiadas; algo que podría evitarse si se contara con alguna herramienta automática que avisara de la imposibilidad de realizar determinados análisis estadísticos o, en una versión más avanzada y ambiciosa, que sugiriera qué técnica sería la más apropiada.

En las secciones anteriores se han sugerido algunas formas que persiguen este mismo objetivo, pero todavía podemos proponer algunas ideas más, que se presentan en esta sección final de la parte metodológica.

Comenzamos con una reflexión sobre la detección de no idoneidad *a posteriori*, algo que ocurre muy frecuentemente en el entrenamiento de RNA. Es decir, a menudo se rechaza una técnica porque el error obtenido es muy elevado. No obstante, conviene tener en cuenta que el error cometido puede ser debido a varios motivos. En primer lugar, el error puede venir condicionado por la existencia de valores perdidos (algo que ya se vio anteriormente) o por la propia distribución del conjunto de los datos (lo que sería un impedimento en el problema). En estos casos, el error suele estar acotado inferiormente, luego cualquier intento de resolución del problema va a ser incapaz de reducir el error que se comete, por venir determinado por la distribución de los datos.

A continuación se van a presentar algunos ejemplos muy sencillos para que se

pueda entender mejor qué queremos decir cuando afirmamos que existen análisis que no se pueden realizar o que a veces no se puede reducir el error que se comete (lo que impide utilizar el error como la única forma de detectar si una técnica se ha aplicado correctamente o no).

Inicialmente, considérese un conjunto de datos bidimensionales que se pueden representar gráficamente en los vértices de un cuadrado (por ejemplo, pensemos en el conjunto $\{(0,0),(2,0),(0,2),(2,2)\}$). Cabe plantearse si, según la dispersión de estos datos, existe una recta que representa mejor que otras la mínima distancia entre dicha recta y los datos (la recta haría las veces de ajuste lineal para esos datos). Sin embargo, si tratamos de calcular una recta de regresión para los datos anteriores, se obtienen dos posibles soluciones: en la notación estadística habitual, se trata de la recta de regresión de X sobre Y y la recta de regresión de Y sobre X. En este ejemplo concreto, las rectas que se obtendrían serían la $x = 1$ y la $y = 1$. Puesto que la covarianza es nula, el índice que se suele utilizar para medir la bondad del ajuste, R^2 sería también cero, lo que se suele interpretar como que estamos explicando el 0 % de la variabilidad del fenómeno analizado o, lo que es lo mismo, que el modelo lineal es muy malo.

Podemos utilizar RNA para ajustar los datos anteriores y la situación no varía excesivamente. Por una parte, es posible conseguir diferentes rectas (normalmente pasarán por el $(1,1)$, pero no necesariamente serán $x = 1$ o $y = 1$). Entrenando varias veces las RNA que ajusten los datos, se comrpueba que las rectas de regresión nos dan un error mínimo, por lo que se deben considerar soluciones óptimas, a pesar de que el error no consigue rebajar un valor bastante poco atractivo.

Supongamos ahora que tenemos un conjunto de datos dispersos, sin ninguna relación aparente entre sus ubicaciones, pero con una característica que los distingue en dos tipos disjuntos (desconocida *a priori*). Pensemos que deseemos realizar una clasificación, es decir, dividir el conjunto de datos en dos conjuntos

disjuntos. Una pregunta lógica es averiguar cuántas divisiones óptimas existen. Según la dispersión de los datos y el procedimiento de división, se puede alcanzar una única solución, infinitas soluciones o ninguna solución. Así, por ejemplo, si se trata de dividir el conjunto por una recta (puede ser interesante utilizar una recta de mínimas distancias que separe el conjunto de datos en dos conjuntos disjuntos), existe al menos una solución si la configuración de los datos es linealmente separable; en cambio, cuando existe un conjunto de datos no linealmente separable, se debe buscar otro procedimiento de separación (por ejemplo, utilizar más de una recta y crear más de dos subconjuntos disjuntos en el plano).

En esta línea, consideremos a continuación un triángulo equilátero de datos; en concreto, pongamos que los tres puntos serían $(0,0)$, $(1,\sqrt{3})$ y $(2,0)$. Como decíamos, desconocemos de qué tipo es cada dato (de las dos posibilidades existentes). Supongamos que tenemos que dividir dicho conjunto de puntos, pero que solo queremos utilizar una recta (que determinaremos gracias a una RNA o mediante otros procedimientos). Sin conocer la clasificación de los 3 datos, ¿cómo creemos que sería la supuesta recta separadora (si buscamos una única solución, probablemente se trataría de una de mínima distancia), que divide en dos conjuntos disjuntos los datos?

En este caso es bastante intuitivo pensar que sería una recta que pasara por un punto definido como el punto más cercano a los tres vértices a la vez, es decir, que pasara por el baricentro, el ortocentro o el circuncentro del triángulo, dependiendo la definición de mínimo que establezcamos; en este triángulo en particular, todos estos puntos coinciden en uno solo, luego no se puede definir la llamada recta de Euler, que sería la que pasa por los tres puntos anteriores. Este hecho nos impide que en este caso en particular se pueda dibujar una única recta que mejor divida el conjunto en dos partes (en el supuesto de no ser un triángulo equilátero, siempre existe la recta de Euler y es una buena candidata para dividir el subconjunto en

dos).

Cada RNA que se entrene puede proporcionar diferentes soluciones (en nuestro caso, todas pasan por el citado baricentro), pero su estimación del error es también poco atractiva, en todos los casos: un 66 %, que coincide con el porcentaje de puntos que pueden resultar mal clasificados al dividir el plano por la recta calculada. Cuando repetimos el ejercicio con RNA simples de clasificación, el error que se comete es siempre el mismo a pesar de que tras cada entrenamiento se obtenga una clasificación distinta. Este error sí podría reducirse incrementando el número de neuronas (y de rectas que separan los datos), pero el problema de la clasificación seguiría deduciéndose de la incapacidad de controlar el tipo de cada dato, por lo que conseguir un subgrupo distinto para cada dato tampoco resolvería el problema de clasificación. En este caso, la dificultad surge porque el procedimiento empleado no proporcionará nunca la salida deseada.

Volviendo al caso del cuadrado, si ahora tratamos de separar los datos sin información *a priori*, se observa que en esta clasificación siempre va existir un error del 50 % (entendido como el porcentaje máximo de puntos que pueden estar incorrectamente clasificados) y que dicho porcentaje no puede rebajarse sin incorporar más neuronas. Supongamos ahora que el conjunto de datos está distribuido en los vértices de un polígono regular con más de 4 caras; en este supuesto, si se desea realizar una clasificación en dos subconjuntos mediante una sola recta que divida el conjunto, se podrán encontrar infinitas soluciones y puede que ninguna de ellas sea mejor que las demás. Si lo que se pretende es obtener una solución relativamente estable, las RNA de clasificación no supervisadas no serían una buena elección en estos casos. Además, como hemos visto antes, el error cometido depende más del conjunto de datos que de la solución encontrada.

Como consecuencia de lo anteriormente expuesto, consideramos comprobado que siempre existirán distribuciones de datos que, por su propia dispersión, siem-

pre producirán un error elevado a la hora de realizar cualquier tipo de análisis de datos. Consideramos que este hecho tiene interés, pues explica que los investigadores no pueden basarse exclusivamente en el análisis del error para validar una metodología y justifica nuestros intentos (pasados y futuros) de buscar herramientas (basadas en RNA) que sean capaces de sugerir técnicas apropiadas para un análisis estadístico adecuado de los datos procedentes de una situación real.

Parte II

APLICACIÓN

Capítulo 4

Presentación del problema

La aplicación que proponemos en esta tesis surge de un problema que tiene gran importancia en la actualidad y que trataremos de ayudar a conseguir una solución. Se trata de entender mejor cuál es la relación entre Educación y Economía. En una primera aproximación, resulta obvio que la Educación tiene influencia real y en la Economía, tanto individual como colectivamente. Además, esta relación es positiva, en el sentido de que la inversión en Educación tiene una influencia beneficiosa para la Economía. Sin embargo, la Educación también depende de la Economía y una Educación suficiente para promover mejoras económicas necesita de un nivel económico suficiente, por encima de un cierto umbral. Por otra parte, rara vez la Economía se supedita a conseguir unos resultados aceptables en Educación. Por tanto, el problema está servido: la Economía necesita de la Educación, la Educación necesita de la Economía, la Economía no se subordina a la Educación ¿y la Educación acaso debería subordinarse a la Economía?

En las últimas décadas las Universidades Españolas se han enfrentado a un reto considerable, porque se le ha permitido el acceso a un número mayor de candidatos a la Educación Superior, suponiendo esto un coste económico mayor

para la Educación en España; pero, al mismo tiempo, la formación que se propone para los estudiantes universitarios cada vez está más dirigida hacia un posterior rendimiento económico, tanto individual como inserto en la sociedad a la que dicho estudiante pertenecerá.

Al mismo tiempo, se ha generalizado el fracaso escolar en diferentes niveles educativos y los informes internacionales sitúan a la educación española (y, más aún, la andaluza) muy por debajo de lo deseado. Al existir una tasa de éxito inferior a lo esperado, sobre todo en el primer año de carrera, se han multiplicado los análisis para localizar las causas y para proponer posibles soluciones.

Otro aspecto a tener en cuenta es el coste de las becas universitarias, en particular cuando se trata de estudiantes de primer año de carrera a quienes se concede beca y no logran los resultados esperados, desperdiciando, desde el punto de vista económico, una gran cantidad de dinero público si el estudiante no termina el curso con éxito.

Finalmente, consideramos pertinente hablar de la influencia de las clasificaciones internacionales de las universidades. La consecución de una adecuada financiación por parte de una universidad, hoy día, depende principalmente de su prestigio (docente, pero sobre todo investigador) y dicho prestigio se alimenta de las buenas posiciones en las clasificaciones más reconocidas. Así, todas las universidades que se preocupen por su crecimiento o supervivencia deben estar atento a cuáles son los criterios académicos (o supuestamente académicos, algunas veces) que les permitirán obtener una financiación suficiente para mantener su nivel de calidad. Indirectamente, este suceso ha obligado a las universidades a mejorar y optimizar el sistema de selección de los estudiantes, para con ello intentar obtener un rendimiento adecuado de los estudiantes que llegan a matricularse en sus distintas titulaciones ofertadas. Simultáneamente, las universidades públicas no pueden perder de vista su sentido social y seguir ofreciendo formación de calidad a

estudiantes con escased de recursos (económicos, culturales y, por qué no decirlo, intelectuales). Sin embargo, no es sencillo establecer una relación entre las características del alumnado y el rendimiento académico que presentará una vez cursada su titulación; incluso menos clara es la posibilidad de establecer una relación entre los estudios superados y la posterior influencia en la economía individual o social.

Estas circunstancias nos sugieren realizar un análisis exhaustivo del comportamiento de los estudiantes, de las variables que afectan a su rendimiento y de los resultados que las Facultades pueden conseguir a nivel global. Aunque ya existen numerosos estudios en España con objetivos parecidos, hasta ahora todos se han visto limitados por la gran cantidad de variables relevantes, por la difícil relación entre ellas, por la dificultad de conseguir una base de datos apropiada, completa y fiable y, finalmente, por la necesidad de utilizar una metodología apropiada y suficientemente flexible como para servir más de ayuda que de dificultad añadida.

4.1. Análisis del problema por parte de otros autores

El interés por el estudio del rendimiento académico de los estudiantes universitarios viene influido por una buena elección de la carrera universitaria, ya que un estudiante que elige con criterio tiene una mayor probabilidad de acabar con éxito, tanto en su vida académica como en su vida profesional, lográndose una mayor rentabilidad social y económica. De hecho, distintos autores (como [4], [8] y [119]) se reafirman en la rentabilidad social que genera una educación universitaria adecuada.

Por otra parte, en general, una mejora en la tasa de éxito de los estudiantes repercutiría en una disminución en los costes de las universidades y, con ello, en un ahorro económico a escala regional o nacional con los estudios universitarios [14].

Es decir, las implicaciones de una buena elección de la carrera académica (a la hora de la matrícula universitaria) y la profesional van más allá del significado de una inversión en educación y, por ello, este factor (elegir correctamente la carrera) es muy influyente tanto en la económica universitaria como en diferentes aspectos considerados y analizados por la Economía de la Educación.

Lógicamente, deducimos que es muy importante detectar los factores que explican el éxito o fracaso de los estudiantes, sobre todo en el primer curso académico (porque el primer año parece que puede marcar el resto de la carrera y, además, porque es sobre el que se puede actuar más directamente mediante una modificación en la elección de la carrera, por ejemplo.

Son numerosos los trabajos que se han desarrollado sobre los factores que influyen en Educación; por ejemplo, se pueden consultar: [10], [88], [138] y [28]. En particular, varios de ellos analizan a estudiantes que estudian Matemáticas de Economía o Empresa; el motivo principal es que estas asignaturas son decisivas para el éxito académico y afectan como ninguna otra en el abandono académico prematuro.

En lo que respecta a las variables previas al ingreso en la Universidad, distintos autores han publicado diferentes trabajos sobre las asignaturas que el estudiante debería haber cursado antes de ingresar; en particular, las asignaturas de Matemáticas resultan esenciales para estudiantes del Grado en Economía, del Grado en Administración y Dirección de Empresas, etc. (cualquiera que sea la institución analizada) [28].

Creemos que resulta relevante resumir cuáles son las distintas variables utilizadas por otros autores cuando tratan de analizar el éxito o fracaso de los estudiantes universitarios. Tras realizar una revisión bibliográfica extensa, podemos enumerar las variables que consideramos más relevantes:

- El nivel educativo de los padres es una de las variables que detectan la mayoría de los autores (en este sentido se puede consultar, por ejemplo, [15]).
- En términos económicos, en la mayoría de los casos se puede comprobar que afecta la renta familiar (ver [11] y [23]); es decir, el que una familia perciba un renta mayor influye de manera positiva en escolarización universitaria.
- En cambio, existe una relación inversa entre el número de personas en el ámbito familiar y el número de estudiantes universitarios [11].
- La región de residencia también afecta significativamente en el rendimiento académico, siendo, por ello, una de las variables que se estudian con más curiosidad en este trabajo.
- Existen muchos trabajos donde se comprueba que la variable sexo influye en el rendimiento final del estudiante en niveles educativos superiores; en particular, [84] y [121] afirman que, bajo determinadas condiciones, las mujeres tienden a alcanzar niveles educativas más elevados que los hombres.

Capítulo 5

Datos

Los datos utilizados en esta tesis han sido recopilados de distintas fuentes: los datos económicos se han consultado en distintas bases de datos y actualizados a partir del Instituto Nacional de Estadística (INE) [78, 79, 80], del Instituto Geográfico Nacional (IGN)(`www.ign.es`), del Instituto de Estadística y Cartografía de Andalucía (IECA) [77], de los Ayuntamientos de Sevilla y Dos Hermanas, de Correos (`http://www.correos.es/ss/Satellite/site/pagina-buscador_codigos_postales/sidioma=es_ES`), de la herramienta web Google Maps (`https://www.google.es/maps`) y la página (`http://www.codigo-postal.info/sevilla/sevilla/7`).

A partir de la información recopilada, se han desarrollado otras variables económicas definidas *ad hoc* y verificadas por expertos en la materia, atendiendo a las necesidades que el problema que deseamos resolver sugería.

Por otra parte, las variables de carácter educativo se han obtenido gracias a la colaboración de los profesores y coordinadores de las asignaturas del Área de Métodos Cuantitativos del Departamento de Economía, Métodos Cuantitativos e

Historia Económica de la Universidad Pablo de Olavide, de Sevilla (UPO). Los datos de acceso e información previa de los estudiantes se han conseguido gracias al Área de Estudiantes y al Área de Gestión Académica de la misma UPO.

5.1. Datos educativos

Los datos educativos (a los que se acaba de hacer referencia) se refieren a un conjunto de estudiantes de la UPO, en concreto, de la Facultad de Ciencias Empresariales. Estos estudiantes pertenecen a cuatro titulaciones distintas:

- Grado en Administración y Dirección de Empresas (GADE)
- Doble Grado en Administración y Dirección de Empresas y Derecho (GADE-GD)
- Grado en Finanzas y Contabilidad (GFC)
- Doble Grado en Finanzas y Contabilidad y Derecho (GFC-GD)

La información académica de estos estudiantes está delimitada por un factor muy determinado, que es el ser alumnos y alumnas de las asignaturas que imparte el Área de Métodos Cuantitativos del Departamento de Economía, Métodos Cuantitativos e Historia Económica, dentro de las cuatro titulaciones citadas anteriormente.

Para concretar algo más, con la información obtenida de cada estudiante se ha realizado un seguimiento de las cuatro titulaciones, obteniendo información del conjunto de asignaturas que se imparten en cada uno de los grados, que a continuación se detallan por curso, semestre, créditos y tipología en las siguientes tablas: Tabla 5.1, Tabla 5.2, Tabla 5.3 y Tabla 5.4.

Tabla 5.1: Asignaturas de formación básica u obligatoria impartidas por el Área de Métodos Cuantitativos en GADE

Asignaturas	**Curso**	**Semestre**	**Créditos**	**Tipo**
Matemática Empresarial I	1º	1º	6	Formación básica
Matemática Empresarial II	1º	2º	6	Obligatoria
Estadística Empresarial I	1º	2º	6	Formación básica
Matemática Financiera	2º	1º	6	Obligatoria
Estadística Empresarial II	2º	1º	6	Obligatoria
Métodos Estadísticos y Econométricos en la Empresa	2º	2º	6	Obligatoria

Fuente: elaboración propia

A continuación se realiza un breve resumen de las distintas asignaturas, atendiendo a sus criterios de evaluación y al proceso de impartición de las mismas. Todas las asignaturas que se describen a continuación siguen un tipo de modelo de docencia denominado como C1, el cual consiste en impartir la asignatura en dos modalidades: sesiones de Enseñanzas Básicas (EB), que suponen el 50 % de la asignatura, y el otro 50 % en sesiones de Enseñanzas Prácticas y de Desarrollo (EPD).

- Matemática Empresarial I

  Toda la información detallada sobre la asignatura impartida en GADE y GADE-GD se puede consultar en [36], [93] y [95]. En la impartida en GFC y GFC-GD se puede consultar [37], [40] y [43].

  - Evaluación: durante el curso se realizan distintas actividades que se tienen en cuenta en la evaluación de la asignatura. Las pruebas que se

Tabla 5.2: Asignaturas de formación básica u obligatoria impartidas por el Área de Métodos Cuantitativos en GADE-GD

Asignaturas	**Curso**	**Semestre**	**Créditos**	**Tipo**
Matemática Empresarial I	1º	1º	6	Formación básica
Matemática Empresarial II	1º	2º	6	Obligatoria
Matemática Financiera	2º	1º	6	Obligatoria
Estadística Empresarial I	2º	2º	6	Obligatoria
Estadística Empresarial II	3º	1º	6	Obligatoria
Métodos Estadísticos y Econométricos en la Empresa	3º	2º	6	Obligatoria

Fuente: elaboración propia

Tabla 5.3: Asignaturas de formación básica u obligatoria impartidas por el Área de Métodos Cuantitativos en GFC

Asignaturas	**Curso**	**Semestre**	**Créditos**	**Tipo**
Matemática Empresarial I	1º	1º	6	Formación básica
Matemática Financiera	1º	2º	6	Obligatoria
Matemática Empresarial II	2º	1º	6	Obligatoria
Estadística para Finanzas I	2º	1º	6	Formación básica
Estadística para Finanzas II	2º	2º	6	Obligatoria
Métodos Estadísticos y Econométricos en Finanzas	3º	1º	6	Obligatoria

Fuente: elaboración propia

Tabla 5.4: Asignaturas de formación básica u obligatoria impartidas por el Área de Métodos Cuantitativos en GFC-GD

Asignaturas	**Curso**	**Semestre**	**Créditos**	**Tipo**
Matemática Empresarial I	1º	1º	6	Obligatoria
Matemática Financiera	1º	2º	6	Obligatoria
Matemática Empresarial II	2º	1º	6	Obligatoria
Estadística para Finanzas I	2º	2º	6	Obligatoria
Estadística para Finanzas II	3º	1º	6	Obligatoria
Métodos Estadísticos y Econométricos en Finanzas	3º	2º	6	Obligatoria

Fuente: elaboración propia

desarrollan durante el curso evalúan tanto las EB como las EPD.

- **Evaluación de las EPD (evaluación continua):** durante la sesiones de las clases de EPD, se llevan a cabo varias pruebas a modo de evaluación continua del estudiante, es decir, para poder realizar un seguimiento del estudiante.
  - Controles por temas: al finalizar cada tema se realiza una evaluación del estudiante resolviendo diversos ejercicios del tema que corresponde. La puntuación máxima es de 2 puntos, sin existir ningún mínimo necesario para aprobar la asignatura.
  - Pruebas virtuales a través de la plataforma WebCT de la UPO: se realiza un test teórico después de cada tema. La puntuación máxima es de un punto y no existe ningún mínimo establecido necesario para aprobar la asignatura.
  - Pruebas de informática: durante el semestre se realizaban 3 prácticas de informática y eran evaluadas al final de las tres

sesiones de informática. Estas pruebas tienen un valor total de 2 puntos, siendo 1 el valor mínimo establecido que permite tener la oportunidad de aprobar la asignatura (es decir, un estudiante no puede aprobar sin superar este mínimo en algún momento).

- **Evaluación de las EB y EPD (prueba final):** esta prueba tiene lugar al final del semestre, con un peso del 50 % de la nota final de la asignatura; este apartado supone 5 puntos de los 10 que como máximo pueden alcanzarse. Estos 5 puntos se distribuyen en 1,5 puntos de conocimientos teóricos y 3,5 puntos correspondientes a la resolución de distintos problemas prácticos.

- Cursos: a efectos de los cálculos realizados en esta tesis, esta asignatura se ha impartido en los cursos académicos 2009-2010, 2010-2011 y 2011-2012.

- Estadística Empresarial I

Para una información más detallada sobre la asignatura impartida en GADE y GADE-GD se puede consultar [69], [57] y [59].

- Evaluación: la evaluación se basa en la realización de unas series de actividades de forma continua a lo largo del curso.

  - **Evaluación de las EB (prueba final):** las clases de EB se evalúan mediante una prueba escrita compuesta por preguntas teóricas, cuestiones teórico-prácticas y problemas relacionados con el temario explicado en la asignatura. Dicha prueba supone un 50 % de la nota total. Al estudiante se le exige 1,5 puntos de los 5 puntos establecidos en el examen para que tengan la oportunidad de aprobar la asignatura.

- **Evaluación de las EPD (evaluación continua):**
  - Controles periódicos: durante el semestre se realizan diversos controles periódicos y se realiza un trabajo individual que supone el 30 % de la nota. No existe un mínimo en este 30 %.
  - Pruebas de informática: durante el semestre se realizan 3 prácticas de informática y son evaluadas al final de las tres sesiones de informática. Estas pruebas tienen un valor total de 2 puntos, siendo 1 el valor mínimo establecido.

- Cursos: cursos académicos 2009-2010, 2010-2011 y 2011-2012.

- Estadística para Finanzas I

Una información más detallada sobre la asignatura impartida en GFC y GFC-GD se puede consultar en [58] y [60].

- Evaluación: durante todo el semestre se realizan unas series de actividades evaluadoras del estudiante.
  - **Evaluación de las EB (prueba final):** la evaluación de las clases de EB se realiza mediante mediante una prueba escrita compuesta por preguntas teóricas, cuestiones teórico-prácticas y problemas relacionados con el temario explicado en la asignatura. Dicha prueba supone un 50 % de la nota total. Se le exige al estudiante 1,5 puntos de los 5 puntos del examen.
  - **Evaluación de las EPD (evaluación continua):** durante las clases de EPD se realizan distintas pruebas.
    1. Pruebas: se realizan diversas pruebas periódicas y un trabajo individual. Para el estudiante, esta evaluación supone el 30 % de la nota. No existe un mínimo en este 30 %.

2. Pruebas de informática: durante el semestre se realizan 3 prácticas de informática y su evaluación consiste en tres pruebas realizadas al final de las tres sesiones de informática. Estas pruebas tienen un valor total de 2 puntos siendo 1 el valor mínimo establecido.

- Cursos: cursos académicos 2010-2011 y 2011-2012.

- Estadística Empresarial II

  La información más detallada se puede consultar en [70] y [71]; esta asignatura se imparte en GADE y en GADE-GD.

  - Evaluación: la evaluación se basa en la realización de unas series de actividades continuas a lo largo del curso.
    - **Evaluación de las EB (prueba final):** las clases de EB se evalúan mediante una prueba escrita compuesta por preguntas teóricas, cuestiones teórico-prácticas y problemas relacionados con el temario explicado en la asignatura. Dicha prueba supone un 50 % de la nota total. Al estudiante se le exige 1,5 puntos de los 5 establecidos en el examen.
    - **Evaluación de las EPD (evaluación continua):**
      - Controles periódicos: durante el semestre se realizan diversos controles periódicos que suponen el 30 % de la nota. No existe un mínimo en este 30 %.
      - Pruebas de informática: durante el semestre se realizan 3 prácticas de informática y son evaluadas al final de las tres sesiones de informática. Estas pruebas tienen un valor total de 2 puntos, siendo 1 el valor mínimo establecido.
  - Cursos: cursos académicos 2010-2011 y 2011-2012.

- Estadística para Finanzas II

  Para una información más detallada sobre la asignatura impartida en GFC y GFC-GD se puede consultar [61].

  - Evaluación: durante todo el semestre se realizan unas series de actividades evaluadoras del estudiante.
    - **Evaluación de las EB (prueba final):** la evaluación de las clases de EB se realiza mediante una prueba escrita compuesta por preguntas teóricas, cuestiones teórico-prácticas y problemas relacionados con el temario explicado en la asignatura. Dicha prueba supone un 50 % de la nota total. Se le exige al estudiante 1,5 puntos de los 5 del examen.
    - **Evaluación de las EPD (evaluación continua):** durante las clases de EPD también se realizan distintas pruebas.
      1. Controles periódicos: se realizan diversos controles periódicos del trabajo individual. Para el estudiante, esta evaluación supone el 30 % de la nota. No existe un mínimo en este 30 %.
      2. Pruebas de informática: durante el semestre se realizan 3 prácticas de informática y su evaluación consiste en tres pruebas realizadas al final de las tres sesiones de informática. Estas pruebas tienen un valor total de 2 puntos, siendo 1 el valor mínimo establecido.
  - Cursos: curso académico 2011-2012.

- Matemática Empresarial II

  Toda la información detallada sobre la asignatura impartida en GADE y

GADE-GD se puede consultar en [38], [39] y [42]. La impartida en GFC y GFC-GD se puede consultar en [94] y [96].

- Evaluación: durante el curso se realizan distintas actividades que se tienen en cuenta en la evaluación de la asignatura. Las pruebas que se desarrollan durante el curso evalúan las clases de EB y EPD.
  - **Evaluación de las EPD (evaluación continua):** durante la impartición de las clases de EPD, se llevan a cabo varias pruebas, a modo de evaluación continua del estudiante, es decir, para poder realizar un seguimiento del estudiante.
    - Controles por temas: al finalizar cada tema se realiza una evaluación del estudiante, quien debe resolver diversos ejercicios del tema que corresponde. La puntuación máxima es de 2 puntos, sin existir ningún mínimo.
    - Pruebas virtuales: a través de la plataforma WebCT de la UPO, se realiza un test teórico, después de cada tema. La puntuación máxima es de un punto y no existe ningún mínimo establecido.
    - Pruebas de informática: durante el semestre se realizan 3 prácticas de informática y son evaluadas al final de cada una de las tres sesiones de informática. Estas pruebas tienen un valor total de 2 puntos, siendo 1 el valor mínimo establecido.
  - **Evaluación de las EB (prueba final):** esta prueba tiene lugar al final del semestre, con un peso del 50 % de la nota final de la asignatura; supone 5 puntos de los 10 que como máximo puede alcanzar el estudiante. Estos 5 puntos tienen una distribución de 1,5 puntos de conocimientos teóricos frente a 3,5 puntos en los que se evalúa la resolución de distintos problemas.

- Cursos: esta asignatura se ha impartido en los cursos académicos 2009-2010, 2010-2011 y 2011-2012.

- Matemática Financiera

Toda la información detallada sobre la asignatura impartida en GADE y GADE-GD se puede consultar en [30] y [32]. La impartida en GFC y GFC-GD se puede consultar en [29], [31] y [33].

- Evaluación: durante el curso se realizan distintas actividades que se tienen en cuenta en la evaluación de la asignatura. Las pruebas que se desarrollan durante el curso, evalúan las clases de EB o EPD.
  - **Evaluación de las EPD (evaluación continua):** durante la impartición de las clases de EPD, se llevan a cabo varias pruebas, a modo de evaluación continua del estudiante, es decir, para poder realizar un seguimiento del estudiante.
    - Controles por temas; al finalizar cada tema se realiza una evaluación del estudiante, quien debe resolver diversos ejercicios del tema que corresponde. La puntuación máxima es de 2 puntos, sin existir ningún mínimo.
    - Pruebas virtuales: a través de la plataforma WebCT de la UPO, se realiza un test teórico después de cada tema. La puntuación máxima es de un punto y no existe ningún mínimo establecido.
    - Pruebas de informática: durante el semestre se realizan 3 prácticas de informática y son evaluadas al final de cada una de las tres sesiones de informática. Estas pruebas tienen un valor total de 2 puntos, siendo 1 el valor mínimo establecido.

- **Evaluación de las EB (prueba final):** esta prueba tiene lugar al final del semestre, con un peso del 50 % de la nota final de la asignatura; supone 5 puntos de los 10 que, como máximo, pueden alcanzarse. Estos 5 puntos tienen una distribución de 1,5 puntos de conocimientos teóricos y 3,5 puntos que engloban la resolución de distintos problemas.

- Cursos: esta asignatura se ha impartido en los cursos académicos 2009-2010, 2010-2011 y 2011-2012.

- Métodos Estadísticos y Econométricos en la Empresa

Toda la información se puede consultar en [104] y [105]; esta asignatura se imparte en GADE y en GADE-GD.

- Evaluación: durante el curso se realizan distintas actividades que se tienen en cuenta en la evaluación de la asignatura. Las pruebas que se desarrollan durante el curso evalúan las clases de EB o EPD.

  - **Evaluación de las EPD (evaluación continua):** durante el desarrollo de las clases de EPD se llevan a cabo varias pruebas, a modo de evaluación continua del estudiante, es decir, para poder realizar un seguimiento del rendimiento del estudiante a lo largo del semestre.

    - Controles por temas: al finalizar cada tema se realiza una evaluación del estudiante, quien debe resolver diversos ejercicios del tema que corresponde. La puntuación máxima es de 2 puntos, sin existir ningún mínimo exigible.

    - Pruebas virtuales: a través de la plataforma WebCT de la UPO, se realiza un test teórico después de cada tema; la pun-

tuación máxima es de un punto y no existe ningún mínimo establecido.

◇ Pruebas de informática: durante el semestre se realizan 3 prácticas de informática y son evaluadas al final de las tres sesiones de informática. Estas pruebas tienen un valor total de 2 puntos, siendo 1 el valor mínimo establecido.

- **Evaluación de las EB (prueba final):** esta prueba tiene lugar al final del semestre, con un peso del 50 % de la nota final de la asignatura; supone 5 puntos de los 10 que pueden alcanzarse como máximo. Estos 5 puntos tienen una distribución de 1,5 puntos de conocimientos teóricos y 3,5 puntos sobre la resolución de distintos problemas.

- Cursos: esta asignatura se ha impartido en los cursos académicos 2010-2011 y 2011-2012.

- Métodos Estadísticos y Econométricos en Finanzas

Para una información más detallada sobre la asignatura impartida en GFC y GFC-GD, se puede consultar [106].

- Evaluación: durante el curso se realizan distintas actividades que se tienen en cuenta en la evaluación de la asignatura. Dichas pruebas evalúan las clases de EB y las de EPD.

  - **Evaluación de las EPD (evaluación continua):** durante la impartición de las clases de EPD, se llevan a cabo varias pruebas, como evaluación continua del estudiante, es decir para poder seguir un seguimiento del rendimiento del estudiante a lo largo del semestre.

- ◇ Controles por temas: al finalizar cada tema se realiza una evaluación del estudiante, quien debe resolver diversos ejercicios del tema que corresponda. La puntuación máxima es de 2 puntos, sin existir ningún mínimo.

- ◇ Pruebas virtuales: a través de la plataforma WebCT de la Universidad Pablo de Olavide, se realiza un test teórico después de cada tema. La puntuación máxima es de un punto y no existe ningún mínimo establecido.

- ◇ Pruebas de informática: durante el semestre se realizan 3 prácticas de informática y son evaluadas al final de las tres sesiones de informática. Estas pruebas tienen un valor total de 2 puntos, siendo 1 el valor mínimo establecido como condición necesaria para poder aprobar la asignatura.

- ○ **Evaluación de las EB (prueba final):** esta prueba tiene lugar al final del semestre, con un peso del 50 % de la nota final de la asignatura; es decir, supone 5 puntos de los 10 que, como máximo, se pueden alcanzar. Estos 5 puntos tienen una distribución de 1,5 puntos de conocimientos teóricos y 3,5 puntos correspondientes a la resolución de distintos problemas prácticos.

- Curso: esta asignatura se ha impartido en el curso académico 2011-2012.

Asignatura	Titulación
Matemática Empresarial I	GADE, GADE-GD, GFC y GFC-GD
Matemática Empresarial II	GADE, GADE-GD, GFC y GFC-GD
Estadística Empresarial I	GADE y GADE-GD
Estadística para Finanzas I	GFC y GFC-GD
Estadística Empresarial II	GADE y GADE-GD
Estadística para Finanzas II	GFC y GFC-GD
Matemática Financiera	GADE, GADE-GD, GFC y GFC-GD
Métodos Estadísticos y Econométricos en la Empresa	GADE y GADE-GD
Métodos Estadísticos y Econométricos en Finanzas	GFC y GFC-GD

Las variables de ámbito educativo que conocemos de los estudiantes se citan a continuación:

- Titulación **[Tit_año]**: información de la titulación del estudiante en el momento en que comenzó a cursar la asignatura. Téngase en cuenta que si un estudiante ha cambiado de titulación, se tiene en cuenta el momento en que cursó la asignatura, no cuando convalidó.

- Curso académico **[Curso]**: información relativa al curso académico en el que el estudiante se matriculó en la asignatura; si ha cursado más de una vez la asignatura, la información se recopila por año académico.

- Número de convocatoria **[Conv.]**: número de convocatorias que el estudiante lleva agotadas en el momento del estudio; el término *agotada* puede referirse a que se haya presentado al examen (lo haya superado o no) o a que no se presentara al examen, ya que las convocatorias se agotan en la UPO independientemente de que el estudiante haya utilizado su derecho a examen. El valor está acotado en $\{1, ..., 6\}$.

- Matrículas **[Curso]**: información referente al número de veces que se ha matriculado en la asignatura, aunque aquí se tiene en cuenta si ha cursado la misma asignatura en otra titulación (si ha cambiado de titulación o ha trasladado expediente). El valor está acotado en $\{1, 2, 3\}$.

- Nota definitiva **[Nota]**: se refiere a la nota (máxima) definitiva obtenida al final de la asignatura, coincidiendo con la nota la máxima alcanzada de entre todas las convocatorias presentadas. Este valor esta comprendido en $[0, 10]$.

- Nota evaluación **[Evaluación]**: nota media de las distintas variables creadas sobre las evaluaciones continuas del estudiante. En este subapartado se definen un conjunto de variables que establecen las distintas evaluaciones continua establecidas. Por ejemplo, en Matemática Empresarial I, los estudiantes realizan tres tipos de pruebas distintas cada semestre; en este caso, se crearían tres variables que cada una representa la media obtenida en cada tipo de prueba. Se debe tener en cuenta que, además, se recopilan las notas de las diferentes pruebas (por separado) que se realizan en cada asignatura.

- Nota final **[Nota final]**: se considera nota final a la nota que alcanza el estudiante una vez aprobada la asignatura; los valores que corresponden a esta variable está en $[5, 10]$.

- Nota de informática **[Nota_informática]**: esta nota es la que alcanza el estudiante una vez realizadas las pruebas que evalúan la informática, ya sea durante la evaluación continua o en el examen final de la asignatura.

- Grupo en clase de EPD **[EPD]**: nombre del grupo de clase de EPD; en esta variable se registra la información sobre el grupo de EPD donde se ha matriculado el estudiante por primera vez; se debe tener en cuenta que este valor no siempre sería constante, ya que a veces se realizan permutas

entre los estudiantes (cambios de grupo); estas permutas se han tenido en cuenta siempre que la información estuviera detallada en los registros de las asignaturas correspondientes.

- Línea **[Línea]**: información sobre la línea (grupo de EB) donde se matriculó cada estudiante por primera vez en cada asignatura; es decir, corresponde a la asignación realizada en el momento de la matrícula; nótese que esta línea puede ser diferente para cada asignatura en un mismo año académico.

Además de las variables anteriores, que determinan las asignaturas, se conocen las siguientes características personales y educativas previas de los estudiantes:

- Sexo **[Sexo]**: variable que toma los valores *mujer* y *hombre*, con un único valor posible para cada estudiante si bien puede haber datos perdidos.
- Edad **[Edad]**: fecha de nacimiento del estudiante.
- Edad2 **[Edad2]**: valor numérico creado a partir de la anterior variable edad, siendo este nuevo valor el correspondiente a la edad del estudiante en el momento de entrar por primera vez en los estudios universitarios.
- Municipio familiar **[Mun_f]**: municipio familiar del estudiante durante el primer curso de estudios universitarios.
- Municipio durante el curso **[Mun_c]**: municipio del estudiante durante el curso, que no tiene que coincidir obligatoriamente con el municipio familiar.
- Provincia familiar **[Prov_f]**: provincia correspondiente a la familia del estudiante durante el primer curso de universidad.
- Provincia durante el curso **[Prov_c]**: provincia de residencia del estudiante durante el curso.

- Dirección postal familiar **[Dir_f]**: dirección de la familia del estudiante durante el primer curso de universidad.
- Dirección postal durante el curso **[Dir_c]**: dirección del estudiante durante el curso, siendo esta variables en algunos casos coincidente con la dirección postal familiar.
- Provincia del centro donde cursó los estudios previos **[Prov_centro]**: provincia del centro de los estudios justo anteriores al acceso a la UPO.
- Municipio del centro donde cursó los estudios previos **[Mun_centro]**: municipio del centro de los estudios justo anteriores al acceso a la UPO.
- Vía de acceso a la Universidad **[Vía_acceso]**: puede tomar los siguientes valores:

  1. Ciclo Formativo (FPII)
  2. Prueba de Acceso a la Universidad (PAU) 2008/2009
  3. Prueba de Acceso a la Universidad (PAU) 2009/2010
  4. Prueba de Acceso a la Universidad (PAU) 2010/2011

- Tipo de centro **[Tipo_centro]**: tipo de centro donde el estudiante ha cursado sus estudios justamente anteriores al acceso a la UPO. Puede tomar los siguientes valores:

  1. I.E.S
  2. C.D.P
  3. Otros

- Nota del expediente Bachillerato **[Nota_exp]**: nota del expediente académico del estudiante en bachillerato; esta nota está comprendida entre 5 y 10.

- Nota fase general de la PAU **[Nota_GPAU]**: nota de la fase general de la PAU (esta nota está calculada como la media obtenida a partir de los exámenes de: Comentario de Texto, Idioma, Filosofía e Historia); el valor de esta nota está comprendido entre 0 y 10.
- Nota acceso **[Nota_Acceso]**: nota general de selectividad, con valor comprendido entre 4 y 14.
- Nota definitiva de acceso a la Universidad **[Nota_def]**: nota general de selectividad, con valor comprendido entre 5 y 14.

5.2. Características socio-económicas

Una vez realizados los cálculos y las modificaciones oportunas a nuestra base de datos de ámbito económico, las variables recopiladas son las que a continuación se describen. Conviene aclarar que los datos corresponden a los municipios de Andalucía, en concreto a un total de 840 municipios.

- Variables de entorno físico y de medio ambiente:
  - Extensión del municipio **[ExtMun_año]**: kilómetros cuadrados de la superficie de cada término municipal completo en el año 2010. Corresponde a la última información publicada en la base cartográfica numérica a escala 25.000 del IGN.
  - Distancia a la capital provincial **[Dist_año]**: kilómetros de distancia entre cada núcleo principal del municipio y la capital de provincia. La fuente utilizada es el INE,
  - Latitud **[Latitud_año]**: latitud en grados y minutos del municipio donde reside la capitalidad del municipio. Es la información publicada en el año 1999 en la base de datos de cartografía del IGN.

- Latitud modificada **[Latitud_Mod_año]**: latitud en minutos del municipio; es una variable de elaboración propia a partir de la variable Latitud.
- Longitud **[Longitud_año]**: longitud en grados y minutos del municipio donde reside la capitalidad del municipio. Es la información publicada en el año 1999 en la base de datos de cartografía del IGN.
- Longitud modificada **[Longitud_Mod_año]**: longitud en minutos del municipio; es una variable de elaboración propia a partir de la variable Longitud.
- Altitud sobre el nivel del mar **[Altitud_mar_año]**: metros de altitud sobre el nivel del mar de un punto de la entidad singular principal. Es la información publicada del año 1999 en la base de datos de cartografía del IGN.

▪ Variables relacionadas con la población

- Población **[Pob_año]**: número de habitantes (a fecha del 1 de enero del año correspondiente) inscritos en el padrón municipal custodiado por el ayuntamiento del municipio. La información obtenida es desde el año 1996 al año 2011, siendo la fuente de obtención de dicha información el INE.La variable Población ha tenido que ser modificada para algunos análisis (con la ayuda de los Ayuntamientos de Sevilla y Dos Hermanas), para obtener información de la población por distritos de la capital de Sevilla así como para obtener información de los distintos núcleos poblacionales de Dos Hermanas; en particular, para obtener información sobre Montequinto, cuyo conocimiento era crucial para nuestro trabajo.
- Población por sexo: mujeres **[Pob_Mujer_año]**: número de habitantes mujeres, a fecha del 1 de enero del año correspondiente, inscritos en

el padrón municipal custodiado por cada Ayuntamiento. La información obtenida es desde el año 1996 al año 2011. La fuente de obtención es INE.

- Edad media de la población **[Edad_Pob_año]**: promedio en años de la edad del total de la población inscrita en el padrón municipal residente en el municipio correspondiente. La información corresponde al período desde el año 2001 hasta 2010 y ha sido obtenida en el INE.
- Población extranjera **[Pob_Ext_año]**: número de habitantes extranjeros, a fecha del 1 de enero del año correspondiente, inscritos en el padrón municipal desde 1997 a 2011.
- Paro registrado **[Paro_año]**: según la orden de 11 de marzo de 1985 del Ministerio de Trabajo y Seguridad Social, el paro registrado lo componen las demandas de empleo pendientes de satisfacer el último día del mes en la Oficinas de Empelo del INEM, excluyendo las correspondientes a las siguientes situaciones:
  1. demandantes ocupados (prestaciones parciales);
  2. demandantes sin disponibilidad inmediata para el trabajo o situación incompatible;
  3. demandantes que solicitan un empleo de características específicas;
  4. trabajadores eventuales agrarios beneficiarios de subsidio especial.

  Los datos de esta variable son proporcionados por el Servicio Público de Empleo Estatal (SEPE) y están referidos al 31 de marzo de cada año.
- Población del año 2001 por edad **[Pob_edad_2001]**: se refiere al número de habitantes, a fecha del 1 de enero del año correspondiente, inscritos en el padrón municipal custodiado por el Ayuntamiento del municipio, segregado por edad en intervalos variables. La información

obtenida es del año 2001, siendo la fuente de obtención de dicha variable el INE.

- Enseñanza y formación: número de centros de enseñanza por municipio y por nivel educativo. La fuente de información es la Consejería de Educación, Cultura y Deporte.
  - Centros Públicos de Educación Básica **[CPEB_año]**.
  - Centros Públicos de Educación Secundaria **[CPES_año]**.
  - Centros Públicos de Educación Infantil **[CPI_año]**.
  - Centros Públicos de Educación Primaria **[CPP_año]**.
  - Centros Públicos Educación Especial **[CPEE_año]**.
  - Centros Públicos de E.S.O **[CPESO_año]**.
  - Centros Públicos de Programas de Garantía Social **[CPPGS_año]**.
  - Centros Públicos de P.C.P.I **[CPPCPI_año]**.
- Agricultura, ganadería y pesca
  - Superficie de cultivo **[Sup_año]**: se proporciona información sobre la superficie, en hectáreas, que ocupan los distintos usos del suelos, dependiendo del tipo de cultivo. La fuente de información es la Consejería de Medio Ambiente de la Junta de Andalucía.
  - Superficie de cultivo herb **[SupHerb_año]**: se proporciona información, en hectáreas, sobre la superficie que ocupan los distintos usos del suelos. La fuente de información es la Consejería de Medio Ambiente de la Junta de Andalucía.
  - Trabajadores eventuales agrarios subsidiados **[Trab_Sub_año]**: número de trabajadores eventuales por municipios. La fuente de información es la Consejería de Medio Ambiente de la Junta de Andalucía.

- Transportes y comunicación
  - Número de vehículos **[NumVeh_año]**: número de vehículos matriculados, en general. La fuente es la Dirección General de Tráfico.
  - Parque de vehículos: número de vehículos matriculados por tipología. La fuente es la Dirección General de Tráfico.
    - Parque de vehículos: turismos **[Turismos_año]**.
    - Parque de vehículos: motocicletas **[Moto_año]**.
    - Parque de vehículos: furgonetas **[Furgo_año]**.
    - Parque de vehículos: camiones **[Camion_año]**.
    - Parque de vehículos: autobuses **[Auto_año]**.
    - Parque de vehículos: tractores industriales **[Tracto_año]**.
    - Parque de vehículos: ciclomotores **[Ciclo_año]**.
    - Parque de vehículos: otros **[Otros_año]**.
  - Parque de camiones: número de camiones matriculados en el municipio; en particular, el número de camiones viene separado por el peso que pueden transportar. La fuente de información es la Dirección General de Tráfico.
    - Parque de camiones: menos de 1 tonelada **[Camión<1t_año]**.
    - Parque de camiones: de 1 a 3 toneladas **[Camión_1t-3t_año]**.
    - Parque de camiones: de 3 a 5 toneladas **[Camión_3t-5t_año]**.
    - Parque de camiones: de 5 a 7 toneladas **[Camion_5t-7t_año]**.
    - Parque de camiones: de 7 a 10 toneladas **[Camion_7t-10t_año]**.
    - Parque de camiones: más de 10 toneladas **[Camion>10t_año]**.
  - Número de vehículos de gasolina **[VehGasolina_año]**: número de vehículos de gasolina matriculados en el municipio. La fuente de información es la Dirección General de Tráfico.

- Número de vehículos de gas-oil **[VehGasoleo_año]**: número de vehículos de gasoil matriculados en el municipio. La fuente de información es la Dirección General de Tráfico.
- Número de vehículos de otros combustibles **[VehOtros_año]**: número de vehículos de otro tipo de combustible diferente a la gasolina y al gasoil matriculados en el municipio. La fuente de información es la Dirección General de Tráfico.
- Número de turismos: menos de 1200 cm^3 **[Tur<1200_año]**. número de vehículos con menos de 1200 cm^3 de cilindrada matriculados en el municipio. La fuente de información es la Dirección General de Tráfico.
- Número de turismos: de 1200 a 1599 cm^3 **[Tur_1200 - 1599_año]**: número de vehículos entre 1200 cm^3 y 1599 cm^3 de cilindrada matriculados en el municipio. La fuente de información es la Dirección General de Tráfico.
- Número de turismos: de 1600 a 1999 cm^3 **[Tur_1600 - 1999_año]**: número de vehículos entre 1600 cm^3 y 1999 cm^3 de cilindrada matriculados en el municipio. La fuente de información es la Dirección General de Tráfico.
- Número de turismos: más de 2000 cm^3 **[Tur>2000_año]**: número de vehículos de más de 2000 cm^3 de cilindrada matriculados en el municipio. La fuente de información es la Dirección General de Tráfico.
- Líneas de la compañía telefónica en servicio **[LíneasTelf_año]**: número de líneas contratadas en la compañía telefónica a 1 de enero del año correspondiente. La fuente es Telefónica.
- Líneas RDSI en servicio **[LíneasRDSI_año]**: número de líneas contratadas en la compañía telefónica a 1 de enero del año correspondiente. La fuente es Telefónica.

- Líneas ADSL en servicio **[LíneasADSL_año]**: número de líneas contratadas en la compañía telefónica a 1 de enero del año correspondiente. La fuente es Telefónica.

- Actividad financiera y empresarial

  - Establecimientos económicos **[EstEco_año]**: número de establecimientos económicos registrados; comprende todos los establecimientos ubicados en el municipio. Hace referencia a una unidad productora de bienes o servicios que desarrolla una o más actividades de carácter económico o social, bajo la responsabilidad de un titular o empresa, en un local situado en un emplazamiento fijo y permanente. La información recopilada de los años 1998, 2004, 2005 y 2008 procede del IEA, concretamente del Directorio de establecimientos con actividades económicas en Andalucía.
  - Establecimientos turísticos **[EstTur_año]**: número de establecimientos turísticos registrados. La información procede de la Dirección General de Turismo y se refieren al año 1998.
  - Números de hoteles **[Hotel_año]**: número de hoteles registrados a 1 de marzo del año correspondiente. La información es la establecida por el Instituto de Estadística y Cartografía de Andalucía.
  - Oficinas de entidades de crédito **[OfiCred_año]**: número de oficinas de entidades de crédito registrado a 1 de marzo del año correspondiente. La información es la establecida por el Instituto de Estadística y Cartografía de Andalucía.
  - Bancos **[Banco_año]**: número de bancos registrado a 1 de marzo del año correspondiente. La información es la establecida por el Instituto de Estadística y Cartografía de Andalucía.

- Oficinas de cajas de ahorros **[Caja_año]**: número de oficinas de cajas de ahorros registrado a 1 de marzo del año correspondiente. La información es la establecida por el Instituto de Estadística y Cartografía de Andalucía.
- Oficinas de cooperativas de créditos **[OfiCoop_año]**: número de oficinas de cooperativas de crédito registrado. La información es la establecida por el Instituto de Estadística y Cartografía de Andalucía.
- Oficinas de establecimientos financieros **[EstFin_año]**: número de oficinas de establecimientos financieros registrado a 1 de marzo. La información es la establecida por el Instituto de Estadística y Cartografía de Andalucía.
- Número de establecimientos de crédito **[EstCred_año]**: número de establecimientos de crédito registrado. Esta variable de elaboración propia ha sido creada a partir de las distintas variables que corresponden a acreedores de créditos: establecimientos financieros, oficinas cooperativas de crédito, cajas de ahorro y bancos. La base de datos se considera para los años desde 1999 hasta 2010.
- Número de plazas hoteleras **[PlazaHot_año]**: número de plazas hoteleras registrado a 1 de marzo. Información establecida según el Instituto de Estadística y Cartografía de Andalucía.
- IBI de naturaleza urbana: nº inmuebles ocio y hostelería **[IBI_ocio_año]**: número de inmuebles de ocio y hostelería que existen en cada municipio. La fuente de información es la Dirección General del Catastro y corresponde al impuesto sobre bienes inmuebles y bienes de naturaleza urbana, del año 1999 hasta el año 2010.
- IBI de naturaleza urbana: nº inmuebles oficina **[IBI_ofic_año]**: número de inmuebles de oficina que existen en cada municipio, encargados en el trámite y tasación del catastro de las viviendas. La fuente de es la

Dirección General del Catastro y corresponde al impuesto sobre bienes inmuebles y bienes de naturaleza urbana, del año 1999 hasta el año 2010.

- Hacienda

  - IRPF: renta media declarada **[RentaN_año]**: variable calculada a partir de la renta neta declarada y el número de habitantes (es decir, la variable Población). Los datos disponibles son de 1997, 1999, 2000, 2002, 2003, 2004 y 2006; todos ellos están medidos en euros excepto el año 1997 que viene definido en pesetas.
  - IRPF: renta neta declarada media **[RentaN_año]**: La renta neta media se define como el cociente entre la renta neta total declarada y el número de declaraciones. La información que se facilita en esta variable está medida en euros y es la que se obtiene como suma de las rentas declaradas según el tipo de rendimiento: rentas netas del trabajo, rentas netas de actividades empresariales, rentas netas de actividades profesionales y otros tipos de rentas. La fuente utilizada es la Agencia Tributaria y la información se refiere a los años 1999, 2000, 2002, 2003, 2004 y 2006.
  - IRPF: número de declaraciones **[NumDecl_año]**: número de declaraciones registradas. La fuente utilizada es la Agencia Tributaria y la información se refiere a los años 1999, 2000, 2002, 2003, 2004 y 2006.
  - I.B.I. de naturaleza urbana: número de recibos **[NumRec_año]**: número de recibos del I.B.I. Se ha considerado como unidad urbana a todos los inmuebles con una relación de propiedad perfectamente delimitada a efectos fiscales. La variable ha sido elaborada a partir de la información de la Dirección General del Catastro, en concreto de sus estadísticas catastraless. Los años disponibles son desde 1998 hasta

2010.

- I.B.I. de naturaleza urbana: base imponible **[NumBase_año]**: valor de la base imponible, atendiendo a que se ha considerado como unidad urbana a todos los inmuebles con una relación de propiedad perfectamente delimitada a efectos fiscales. La variable ha sido elaborada a partir de la información de la Dirección General del Catastro, en concreto de sus estadísticas catastraless. Los años disponibles son desde 1998 hasta 2010.

- Consumo

  - Energía **[Energía_año]**: datos procedentes de las facturaciones en megavatios por hora realizadas por la Compañía Sevilla de Electricidad a sus abonados. Se debe de tener en cuenta que existen municipios que no poseen suministros, luego allí los datos son estimados. La fuente de la base de datos es la propia Compañía Sevillana de Electricidad, del año 1998 hasta el 2010.
  - Consumo de agua invierno: datos procedentes de las facturaciones en metros cúbicos de los meses de invierno (noviembre a abril) realizadas por la Compañía de Agua a los abonados. Se debe tener en cuenta que existen municipios que no poseen suministros, luego allí los datos son estimados. La fuente de la base de datos es la correspondiente compañía de aguas, del año 1998 hasta el 2010.
  - Consumo de agua verano: datos procedentes de las facturaciones en metros cúbicos de los meses de verano (mayo a octubre) realizadas por la compañía de aguas a los abonados. Se debe tener en cuenta que existen municipios que no poseen suministros, luego los datos son estimados. La fuente de la base de datos es la correspondiente compañía de agua, del año 1998 hasta el 2010.

5.3. Análisis descriptivos

A continuación se presenta un análisis preliminar de los datos a fin de estar familiarizados con sus características más generales antes de aplicarles técnicas específicas.

- **Titulación**

  La distribución de los 1492 estudiantes por titulación puede consultarse en la Tabla 5.5.

Tabla 5.5: Tabla de frecuencia por titulaciones

Titulación	Frecuencia
GADE	459
GADE-GD	356
GFC	498
GFC-GD	179

- **Curso académico**

  Como puede observarse en la Tabla 5.6, los datos del estudio realizado proceden de tres cursos académicos consecutivos.

- **Sexo**

  El número de estudiantes de los que tenemos información sobre su sexo es de 1435 (véase la Tabla 5.7).

Tabla 5.6: Tabla de frecuencia por curso académico

Curso	Frecuencia
2009/2010	499
2010/2011	501
2011/2012	492

Tabla 5.7: Tabla de frecuencia por sexo

Sexo	Frecuencia
Hombre	816
Mujer	619

- **Edad**

  Los estudiantes matriculados en asignaturas cuantitativas de la FCE de la UPO durante el período de estudio tienen edades comprendidas entre los 18 y los 48 años de edad. La información puede consultarse en la Tabla 5.8. Recuérdese que esta variable ha sido modificada teniendo en cuenta la primera vez que se matricula cada estudiante (según se comentó al describir la variable).

- **Tipo de centro donde cursó Bachillerato**

  Los alumnos de la FCE de la UPO provienen de centros con diferentes características, como puede ser que se trate de centros públicos o concertados o privados. Además, hay alumnos que han podido realizar algún cambio de tipo de centro durante sus estudios. En la Tabla 5.9 se resume la información relativa al tipo de centro.

Tabla 5.8: Tabla de frecuencia por edad

Edad	Frec.	Edad	Frec.	Edad	Frec.	Edad	Frec.
18	1	25	44	32	6	40	1
19	223	26	22	33	8	41	3
20	351	27	16	34	1	42	1
21	387	28	12	35	4	43	1
22	127	29	14	37	3	44	5
23	103	30	10	38	2	45	1
24	60	31	7	39	1	48	3

Tabla 5.9: Tabla de frecuencia por tipo de centro

Tipo de centro	Frecuencia
C.D.P	590
I.E.S	516
Otros	98

- **Tipo de acceso**

  Creemos que es conocer las distintas modalidades de acceso a la Universidad. En la Tabla 5.10 se puede consultar dicha información.

Creemos que con la información anterior ya es posible hacerse una idea de la base de datos con la que trabajamos a partir de este momento. A continuación se realizarán diferentes tipos de análisis según los objetivos prácticos que queremos alcanzar.

Tabla 5.10: Tabla de frecuencia por tipo de acceso a la Universidad

Tipo de acceso	Frecuencia
Ciclo Formativo	76
PAU 2008/09	500
PAU 2009/10	338
PAU 2010/11	322

5.4. Sobre la procedencia de los estudiantes

Las primeras decisiones importantes que debe tomar un estudiante que va a matricularse en la universidad es decidir qué titulación desea estudiar y, además, dónde desea hacerlo. Sin embargo, estas dos decisiones vienen determinadas en gran medida por el lugar donde el estudiante vive. Lo que no es tan claro es el nivel de influencia que la proximidad geográfica (entre el domicilio o el lugar donde se cursaron los estudios preuniversitarios y la universidad) ejerce sobre la elección.

En este apartado, aprovecharemos la descripción preliminar de algunas características de los datos (concretamente los que se refieren a la procedencia de los estudiantes que eligen ser alumnos de la UPO), para resolver algunas cuestiones conducentes a la cuantificación de la influencia que la cercanía tiene en la decisión final del alumno universitario para estudiar en la UPO.

En un primer vistazo, observamos que existe una gran diversidad en cuanto a la procedencia de estudiantes que se matriculan en las distintas titulaciones de la Facultad de Ciencias Empresariales (FCE) de la UPO. Sin embargo, no es posible establecer un patrón claro que permita determinar qué variables son las que tienen mayor protagonismo en este fenómeno. En lo que sigue, trataremos de explicar la distribución de los estudiantes por la población del municipio del estudiante y por la proximidad entre su domicilio y la UPO, atendiendo tanto a la distancia puramente kilométrica como al tiempo que emplea cada estudiante en llegar a la Universidad. Aquí, la utilización de las RNA es meramente testimonial, puesto que no se realizan mejoras significativas a las técnicas ya conocidas, pero se abre la posibilidad a su aplicación en futuros retos investigadores.

5.4.1. Descripción preliminar de los datos

En esta sección utilizaremos, en concreto, un total de 1406 estudiantes (de los 1492 de la muestra total con que se cuenta). El número de individuos considerados viene determinado por la disponibilidad de información fiable sobre el domicilio familiar. La información disponible se resume a nivel de códigos postales en la Tabla B.1 del Anexo B.

Los datos de los alumnos corresponden a tres cursos académicos distintos (aunque consecutivos), luego todas las variables vienen afectadas por el año en el que cada estudiante se matricula en la universidad por primera vez, siendo los posibles cursos: 2009/2010, 2010/2011 o 2011/2012.

Las provincias de procedencia de los estudiantes por curso académico aparecen recogidas en la Figura 5.1 y en la Tabla 5.11. En la Figura 5.1 se puede apreciar que la muestra de alumnos no es representativa del conjunto de provincias españolas.

Figura 5.1: Provincias españolas con algún alumno matriculado por primera vez en la FCE de la UPO durante los años 2009-2012

No obstante, en la Tabla 5.11 se puede observar que existe un conjunto signi-

Tabla 5.11: Provincias de las que procede algún alumno (según el año)

CURSO		
2009-2010	2010-2011	2011-2012
Almería	Almería	
Badajoz	Badajoz	Badajoz
	Cáceres	
Cádiz	Cádiz	Cádiz
	Ciudad Real	
Córdoba	Córdoba	Córdoba
Huelva	Huelva	Huelva
Jaén	Jaén	Jaén
	Las Palmas	
	Málaga	Málaga
		Navarra
	Palma	
Sta. Cruz de Tenerife	Sta. Cruz de Tenerife	
Sevilla	Sevilla	Sevilla
	Toledo	
Vigo		
469	460	477

Fuente: elaboración propia

ficativo de provincias que se repiten en todos los cursos académicos.

Si se atiende a la información de la Tabla B.1 del Anexo B, se puede entender mejor la distribución. En general, como era de esperar, la mayor parte de los alumnos procede de Andalucía (1370) y de Badajoz (23). Específicamente, solo 36 alumnos (un 2,56 % del total de 1406 alumnos) procede de fuera de Andalucía. Por este motivo, se ha decidido despreciar esta información (al no ser representativa del conjunto nacional) a la hora de relacionarla con las demás variables que puedan explicar las causas por las que los alumnos eligen estudiar en la FCE de la UPO. También se omitirá a estos alumnos en algunos análisis posteriores en los que se utilicen variables en que puedan tener una significativa influencia las características de la comunidad autónoma a la que se refieran los datos. Sin embargo, creemos conveniente hacer notar aquí que la información de la provincia de Badajoz sí podría ser relevante para futuros estudios relacionados con la UPO.

Atendiendo a las distintas provincias de Andalucía, la distribución queda según los datos de la Tabla 5.12 y su representación en la Figura 5.2.

Figura 5.2: Provincias andaluzas según los alumnos matriculados por primera vez en la FCE de la UPO durante los años 2009-2012

Como se puede apreciar, un $83,90\,\%$ de los estudiantes provienen de la provincia de Sevilla. Esto sugiere una segunda criba para aquellos análisis en los que la

Tabla 5.12: Número de alumnos matriculados por primera vez en la FCE de la UPO en 2009-2012 por provincias andaluzas

Provincia	Nº de estudiantes	Porcentaje
Almería	3	0,20 %
Cádiz	141	10,30 %
Córdoba	22	1,60 %
Huelva	46	3,40 %
Jaén	2	0,10 %
Granada	0	0 %
Málaga	7	0,50 %
Sevilla	1149	83,90 %

Fuente: elaboración propia

variable provincia pueda tener un efecto significativo en otras variables relevantes, puesto que no podemos considerar la muestra de alumnos como representativa a nivel autonómico (de todas las provincias o todos los municipios de Andalucía). Así, por ejemplo, para analizar la influencia de la población o la distancia a la UPO del municipio del domicilio familiar de los alumnos, se ha preferido limitar el estudio a la provincia de Sevilla.

Según los datos recogidos en la Tabla 1 del Anexo I, los estudiantes inicialmente considerados proceden de 217 códigos postales distintos, que corresponden a 156 municipios de toda España; de ellos, 92 códigos postales y 68 municipios son de la provincia de Sevilla. La distribución de estudiantes por municipios de la provincia de Sevilla viene recogida en la Tabla 5.13 y representada en la Figura 5.3; la escala viene determinada por la intensidad del color, siendo el más oscuro el municipio con el mayor número de estudiantes.

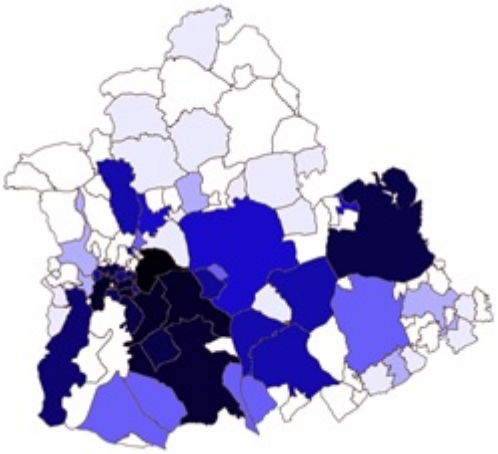
A algunos efectos, sí será posible utilizar la muestra como representativa en

Tabla 5.13: Número de alumnos matriculados por primera vez en la FCE de la UPO en 2009-2012 por municipios sevillanos

Sevilla capital	49,13 %	Paradas	0,09 %
Montequinto	6,35 %	Marchena	0,43 %
Coria del Río	1,22 %	Osuna	0,26 %
Bollullos de la Mitación	0,17 %	El Saucejo	0,09 %
Almensilla	0,09 %	Los Corrales	0,17 %
Gelves	0,61 %	Martín de la Jara	0,09 %
Puebla del Río	0,17 %	Dos Hermanas	6,26 %
Isla Mayor	0,09 %	Utrera	2,52 %
Alcalá del Río	0,35 %	Los Palacios y Villafranca	1,57 %
Guillena	0,43 %	Las Cabezas de San Juan	0,26 %
Burguillos	0,17 %	Lebrija	0,26 %
Castilblanco de los Arroyos	0,09 %	El Coronil	0,26 %
Almadén de la Plata	0,09 %	Montellano	0,26 %
La Rinconada	0,26 %	Sanlúcar la Mayor	0,17 %
Brenes	0,09 %	Olivares	0,61 %
Cantillana	0,17 %	Umbrete	0,43 %
El Pedroso	0,09 %	Espartinas	1,30 %
Cazalla de la Sierra	0,17 %	Villanueva del Ariscal	0,17 %
Guadalcanal	0,09 %	Pilas	0,70 %
Écija	0,87 %	Aznalcázar	0,09 %
Carmona	0,35 %	Villamanrique de la Condesa	0,09 %
La Campana	0,09 %	Camas	0,78 %
Cañada del Rosal	0,35 %	Valencina de la Concepción	0,52 %
Lora del Río	0,09 %	Castilleja de Guzmán	0,17 %
Alcolea del Río	0,09 %	Salteras	0,26 %
Peñaflor	0,09 %	San Juan de Aznalfarache	0,70 %
Alcalá de Guadaíra	4,96 %	Mairena del Aljarafe	3,74 %
Mairena del Alcor	0,52 %	Palomares del Río	0,78 %
El Viso del Alcor	0,26 %	Bormujos	1,74 %
Morón de la Frontera	0,43 %	Tomares	4,26 %
Estepa	0,17 %	Castilleja de la Cuesta	0,87 %
Casariche	0,09 %	Gines	0,78 %
La Roda de Andalucía	0,09 %	Santiponce	0,35 %
Arahal	0,43 %	La Algaba	0,26 %

Fuente: elaboración propia

Figura 5.3: Municipios sevillanos según los alumnos matriculados por primera vez en la FCE de la UPO durante los años 2009-2012



estudios universitarios de la provincia de Sevilla, pero todavía podría parecer que existe alguna distorsión según la proximidad o población de los municipios sevillanos. Por eso, a continuación se diseña una partición de los mismos, atendiendo a la representación gráfica de la Figura 5.3.

Como se puede apreciar en dicho mapa de la provincia de Sevilla (Figura 5.3), la mayor parte de los estudiantes matriculados proviene de los municipios más cercanos a la capital. Parece conveniente realizar una aproximación al concepto de área metropolitana; utilizaremos para ello la definición de *Corona 1* establecida en [22]:

$$Corona1 = Corona1A + Corona1B, \text{ donde}$$

Corona 1A: Alcalá de Guadaíra, La Algaba, Camas, Dos Hermanas, La Rinconada, San Juan de Aznalfarache y Montequinto[1].

[1] Aunque el núcleo poblacional de Montequinto forma parte de Dos Hermanas, nos interesa segregarlo por la relevancia que tiene para la UPO, debido a su proximidad y a su naturaleza de barrio residencial para estudiantes de la propia UPO.

Corona 1B: Castilleja de la Cuesta, Tomares, Mairena de Aljarafe, Gines y Palomares del Río.

En la Tabla 5.14 se presentan los porcentajes de alumnos procedentes de cada municipio de la Corona 1 sobre el total de alumnos de nuestra muestra. Nótese que el 49,13 % de la muestra a estudiar son alumnos de Sevilla capital, pero dicho porcentaje se incrementa hasta un 79,13 % si consideramos Sevilla capital más la Corona metropolitana 1.

Tabla 5.14: Número de alumnos matriculados por primera vez en la FCE de la UPO en 2009-2012 por municipios de la Corona 1

Sevilla capital	49,13 %
Montequinto	6,35 %
La Rinconada	0,26 %
Alcalá de Guadaíra	4,96 %
Dos Hermanas	6,26 %
Camas	0,78 %
San Juan de Aznalfarache	0,70 %
Mairena del Aljarafe	3,74 %
Palomares del Río	0,78 %
Tomares	4,26 %
Castilleja de la Cuesta	0,87 %
Gines	0,78 %
La Algaba	0,26 %

Fuente: elaboración propia

Además de la información presentada anteriormente (por distritos), conviene aclarar que cuando nos centramos en los estudiantes procedentes de Sevilla Capital, podemos realizar los análisis según los códigos postales o según los once distritos de la capital.

5.4.2. Diferentes análisis de los datos

Una vez considerado que los distintos municipios de la provincia de Sevilla podían constituir una muestra interesante de las procedencias de los estudiantes, se decidió analizar la relación entre el número de estudiantes matriculados en la FCE de la UPO y la población (es decir, el número de habitantes) de cada municipio atendiendo al censo del año académico en el que cada estudiante se matriculó por primera vez en la Universidad. Para ello, se define una variable *Población* que afecta a cada estudiante y que se define como el valor medio de las poblaciones anuales del municipio de origen del estudiante, atendiendo a las distintas poblaciones obtenidas al observar la información de cada estudiante por municipio y curso académico. Se detalla en el siguiente ejemplo. Supongamos que un estudiante ingresa en la UPO en el curso 2009/2010, que otro estudiante ingresa en el curso 2010/2011 y que ambos estudiantes provienen del mismo municipio. En este caso, el valor de la población relativo al primer estudiante es el proporcionado en el año 2009, mientras que para el segundo alumno se toma como referencia el valor de la población en el año 2010. Luego, en particular, si no hubiera más esdutiantes procedentes de dicho municipio, la población de este municipio con dos estudiantes, a efecto de nuestros cálculos posteriores, sería la media de los dos valores obtenidos. Téngase en cuenta que se utiliza como valor de referencia el último año donde vive el estudiante antes de entrar en la UPO.

Influencia de la población

Se trata de estimar la influencia que el número de habitantes de un municipio tiene sobre el número de alumnos matriculados en la FCE de la UPO; es decir, la lógica y la probabilidad dictan que los municipios con más habitantes deberían tener más alumnos "representantes" en el análisis, pero es coherente considerar

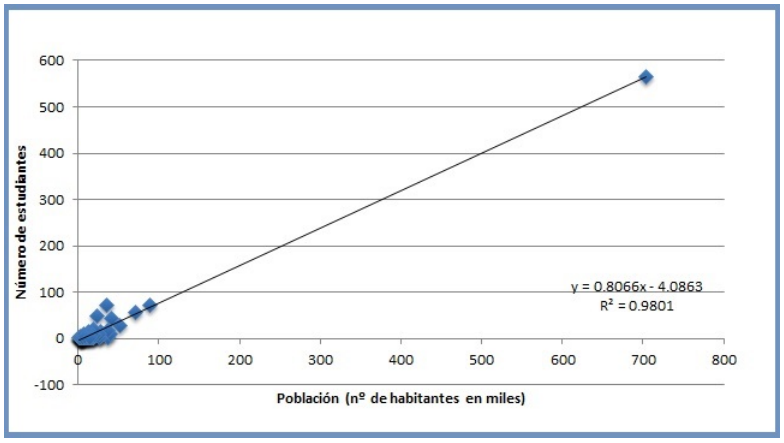
que haya otras variables relevantes aparte de la propia población, así que vamos a tratar de saber si la variable *Población* es suficiente para explicar un grado razonable de variabilidad de la frecuencia en cuanto al número de alumnos de cada municipio que estudian en la FCE de la UPO.

Veamos lo que ocurre según los diferentes grupos de municipios que podamos considerar:

- Provincia de Sevilla

  Analizando la relación que existe entre la población y el número de estudiantes procedentes de cada municipio y atendiendo a los estudiantes de toda la provincia de Sevilla, se puede apreciar una correlación muy fuerte entre estas dos variables (ver la Figura 5.4); de hecho, el modelo explica el 98 % de la variabilidad del número de estudiantes a partir del número de habitantes. Sin embargo, la presencia del *outlier* constituido por la capital puede distorsionar bastante los resultados de la regresión.

Figura 5.4: Relación entre la población de los municipios de la provincia de Sevilla y los estudiantes de la FCE de la UPO

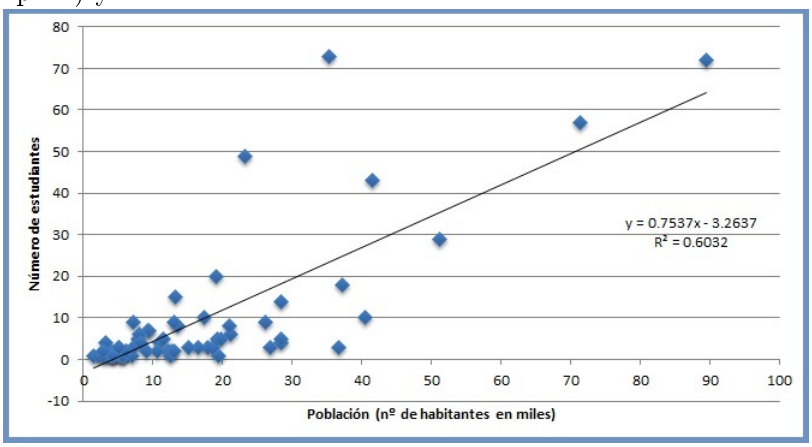


Para realizar un análisis más fiable, volvemos a repetir este cálculo pero omitiendo los estudiantes de la capital y la correspondiente población.

- Provincia de Sevilla menos Sevilla Capital

  Al suprimir la capital, se observa una reducción en la correlación entre las variables. Este hecho sucede por una reducción en el radio de observación entre los datos; se puede observar en la Figura 5.5, tras compararlo con la Figura 5.4.

Figura 5.5: Relación entre la población de los municipios de la provincia de Sevilla (sin la capital) y los estudiantes de la FCE de la UPO



- Sevilla capital más la Corona 1

  Consideremos ahora los datos de la capital junto con los de los municipios más próximos (a la capital, entre sí y a la UPO). Realizamos un análisis descriptivo somero para comprobar si existe una correlación fuerte entre la población y el número de estudiantes. Para que sea más fiable el análisis, dividimos la población de la capital (que representaba casi un 50 % del total, según sus once distritos. Los datos y el resultado de la regresión pueden consultarse en la Tabla 5.15 y en la Figura 5.6.

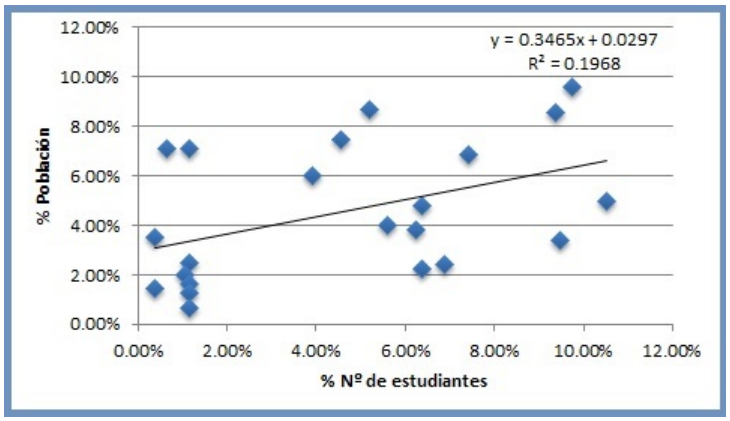
  Una vez establecidos estos cambios lógicos en las variables, llegamos a la conclusión de que no existe una correlación fuerte entre la variable "población" y el número de estudiantes, atendiendo a los distintos distritos y los

Tabla 5.15: Distribución porcentual del nº de estudiantes y la población de los municipios de la Corona 1 y los distritos de Sevilla capital

DTTO.-MUN.	METRO	Nº DE EST.	POBL.	% EST.	% POBL.
Casco Antiguo	Sí	31	60437	3,96 %	5,79 %
Nervión	Sí	83	51578	10,60 %	4,94 %
Cerro Amate	Sí	51	90433	6,51 %	8,67 %
Los Remedios	Sí	60	25038	7,66 %	2,40 %
Montequinto	Sí	73	35374,61	9,32 %	3,39 %
Mairena del Aljarafe	Sí	43	41452.12	5,49 %	3,97 %
Macarena	No	35	77929	4,47 %	7,74 %
Sur	No	9	74027	1,15 %	7,09 %
Norte	No	5	74131	0,64 %	7,10 %
San Pablo-Santa Justa	No	30	62921	3,83 %	6,03 %
Este	No	73	99971	9,32 %	9,58 %
Bellavista-La Palmera	No	49	39719	6,26 %	3,81 %
Alcalá de Guadaíra	No	57	71453,65	7,28 %	6,85 %
Algaba	No	3	15150,67	0,38 %	1,45 %
Camas	No	9	26086	1,15 %	2,50 %
Dos Hermanas	No	72	89533,59	9,20 %	8,58
La Rinconada	No	3	36641	0,38 %	3,51 %
San Juan de Aznalfarache	Sí	8	21026	1,02 %	2,02 %
Castilleja de la Cuesta	No	10	17282	1,28 %	1,66 %
Tomares	No	49	23270,67	6,26 %	2,23 %
Gines	No	9	13108	1,15 %	1,26 %
Palomares del Río	No	9	7143,44	1,15 %	0,68 %

Fuente: elaboración propia

Figura 5.6: Relación entre la población de Sevilla capital y los municipios de la Corona 1 y los estudiantes de la FCE de la UPO

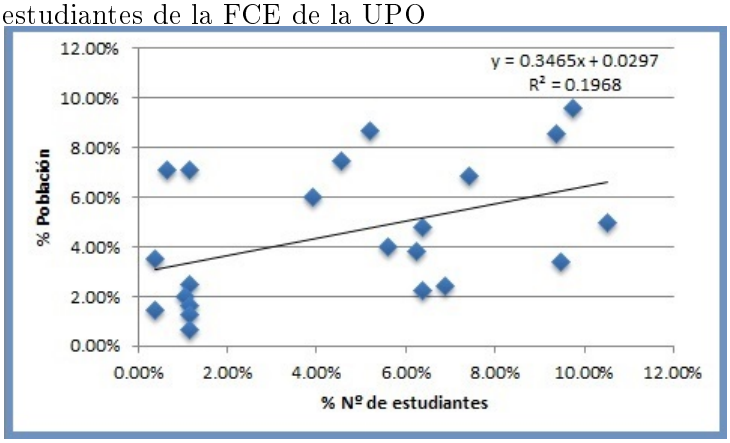


municipios de la Corona 1. Finalmente, vamos a intentar un análisis similar pero eliminando los municipios distintos de la capital.

- Sevilla capital, según sus distritos

  Si nos centramos solo en los once distritos de Sevilla, podemos observar que la correlación de las variables es cada vez menor, aunque este resultado puede venir afectado por la reducción del número de estudiantes y del radio de observación (ver la Tabla 5.16 y en la Figura 5.7).

Figura 5.7: Relación entre la población de los distritos de Sevilla capital y el número de estudiantes de la FCE de la UPO



- Conclusión

  Una vez concluido este primer análisis preliminar de la variable *Población*, se puede concluir que el número de estudiantes matriculados en la FCE de la UPO no viene únicamente explicado por el número de habitantes de cada municipio, pues al menos depende del radio de observación. En lo que sigue trataremos de incorporar al estudio una nueva variable cuya influencia también parece lógica: la distancia a la UPO del domicilio de cada estudiante.

Influencia de la distancia (en km) a la UPO

En un primer intento de valorar la influencia de la ubicación del domicilio familiar sobre la elección de la carrera universitaria, veamos a continuación la relación que existe entre la distancia del municipio del estudiante a la UPO y el número de alumnos de la FCE de la UPO que provienen de cada municipio. Esta relación se ha analizado teniendo en cuenta dos distancias distintas (del municipio a la UPO):

Tabla 5.16: Distribución porcentual del nº de estudiante y la población de los municipios de la Corona 1 y los distritos de Sevilla capital

DTTO.-MUN.	METRO	Nº DE EST.	POBL.	% EST.	% POBL.
Casco Antiguo	Sí	31	60437	6,61 %	8,56 %
Nervión	Sí	83	51578	17,70 %	7,30 %
Cerro Amate	Sí	51	90433	10,87 %	12,80 %
Los Remedios	Sí	60	25038	12,79 %	3,54 %
Macarena	No	35	77929	7,46 %	11,03 %
Sur	No	9	74027	1,92 %	10,48 %
Norte	No	5	74131	1,07 %	10,49 %
San Pablo-Santa Justa	No	30	62921	6,40 %	8,91 %
Eeste	No	73	99971	15,57 %	14,15 %
Bellavista-La Palmera	No	49	39719	10,45 %	5,62 %
Triana	No	43	50181	9,17 %	7,10 %

Fuente: elaboración propia

Distancia 1: distancia media de los estudiantes domiciliados en cada municipio.

Distancia 2: distancia al centro del municipio del estudiante.

Es importante tener claro en cada momento qué tipo de distancia se está utilizando para poder entender la diferencia que existe entre los distintos resultados obtenidos.

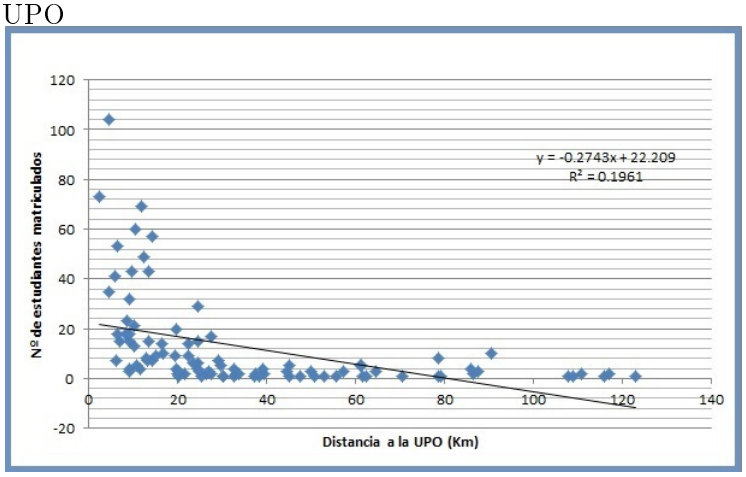
- Provincia de Sevilla

  Centrándonos en la provincia de Sevilla y tomando la distancia de cada municipio (distancia 2), se puede apreciar que existe una correlación inversa, aunque muy débil, entre el número de estudiantes del municipio y la distancia a la UPO desde el municipio familiar (ver la Figura 5.8).

- Provincia de Sevilla sin la Capital

  Observamos que la correlación es más débil que aún con toda la provincia de Sevilla; apenas se puede afirmar que exista una correlación entre las

Figura 5.8: Relación entre la distancia a la Universidad y los estudiantes de la FCE de la UPO



variables (ver la Figura 5.9).

- Sevilla más la Corona 1

  En cambio, cuando realizamos una comparativa entre los estudiantes que proceden de la Corona 1 o de Sevilla Capital distribuida por distritos, se observa que la correlación inversa que existe entre estas dos variables es mucho mayor que la establecida en la provincia de Sevilla. Es decir, que la distancia viene correlacionada con el número de estudiantes próximos a la capital (ver la Figura 5.10).

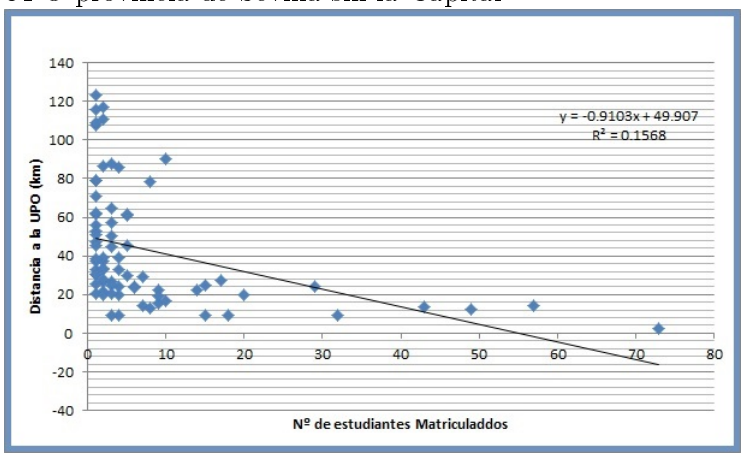
- Sevilla capital por distritos

  Si nos centramos en los once distritos de Sevilla, la correlación que existía se pierde, probablemente por tratarse de un conjunto muy reducido de alumnos (ver la Figura 5.11).

- Sevilla capital por Código Postal

  Considerado el código postal donde se ubica el municipio, se utiliza aquí la distancia al centro de la zona con dicho código postal. En la Figura 5.12, se

Figura 5.9: Relación entre la distancia a la Universidad y los estudiantes de la FCE de la UPO provincia de Sevilla sin la Capital



puede ver que existe una correlación entre estas dos variables que es superior a la establecida por distritos (posiblemente porque se trata de 20 códigos postales).

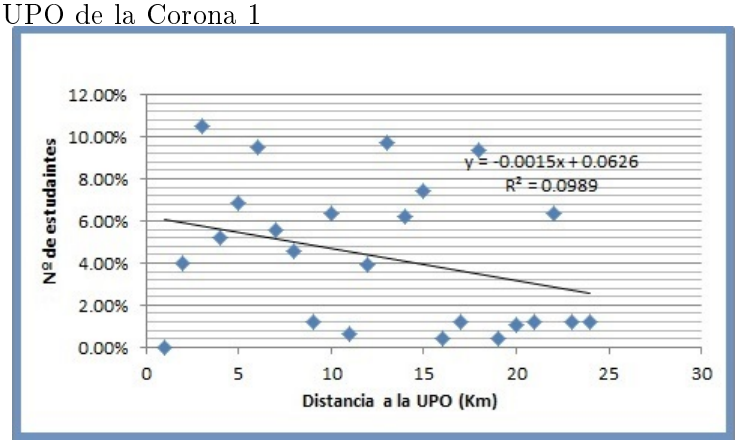
- Conclusión

  Según lo que se ha visto en este apartado, en el área metropolitana formada por la Corona 1 y Sevilla Capital, la correlación que existe entre la variable distancia y las variables del número de alumnos en la FCE de la UPO es inversa y débil. El que la relación sea inversa era esperable, pero el que sea tan débil nos hace pensar en buscar otras explicaciones adicionales para entender la distribución de la población estudiada.

Influencia de la distancia (en segundos) a la UPO

En vista de que la distancia en kilómetros entre el domicilio familiar y la UPO no parece explicar tampoco la distribución de alumnos en la FCE de la UPO, trataremos de utilizar otra variable distancia: la del tiempo medio que se tarda

Figura 5.10: Relación entre la distancia a la Universidad y los estudiantes de la FCE de la UPO de la Corona 1



en llegar del domicilio familiar a la UPO. Para ello, consideraremos como variable *proxy* el tiempo que se tarda en el desplazamiento en un día con tráfico fluido y en vehículo privado[2].

Obviamente, hay una correlación muy fuerte entre la distancia en kilómetros y la distancia en segundos; de hecho, es de 0,97 en los municipios considerados en este estudio. Sin embargo, consideramos de interés valorar si la variable "tiempo" aporta algo más de información a la hora de explicar el número de estudiantes de la FCE de la UPO por municipio (o distrito). Aquí también se han tomado dos definiciones distintas para medir el tiempo:

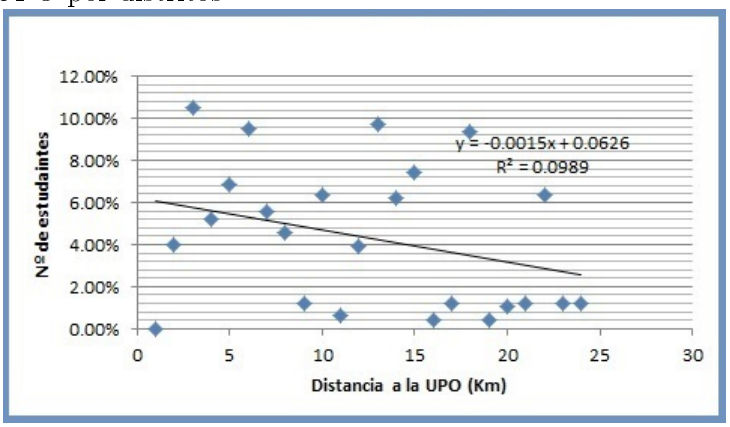
Tiempo 1: viene determinado por el tiempo medio de los estudiantes de ese municipio.

Tiempo 2: es el empleado desde el centro del municipio a la UPO.

Y para el caso de los códigos postales se ha utilizado el tiempo empleado desde el centro de la zona con dicho código postal a la UPO.

[2]Fuente: Google Maps, `http://www.google.es/maps`.

Figura 5.11: Relación entre la distancia a la Universidad y los estudiantes de la FCE de la UPO por distritos



- Provincia de Sevilla

  En el caso del conjunto de los municipios de la provincia de Sevilla, se puede observar que la correlación entre la variable número de alumnos procedentes de cada municipio y el tiempo medio en llegar a la UPO desde el municipio es inversa y muy débil (ver la Figura 5.13).

- Provincia de Sevilla sin la Capital

  En este caso, al reducir el radio de observación, la correlación es más débil que la que existía anteriormente (ver la Figura 5.14).

- Sevilla más la Corona 1

  Si nos centramos en Sevilla Capital y los municipios de la Corona 1, se observa que aumenta la correlación que existe entre el tiempo y el número de alumnos, pero la relación sigue siendo débil e inversa (ver la Figura 5.15).

- Sevilla Capital por distritos

  En cambio, al considerar los datos según los once distritos de Sevilla, la correlación que existía se diluye, probablemente por tratarse de conjuntos reducidos de alumnos, como ocurría con la distancia en kilómetros (ver la

Figura 5.12: Relación entre la distancia a la Universidad y los estudiantes de la FCE de la UPO por código postal

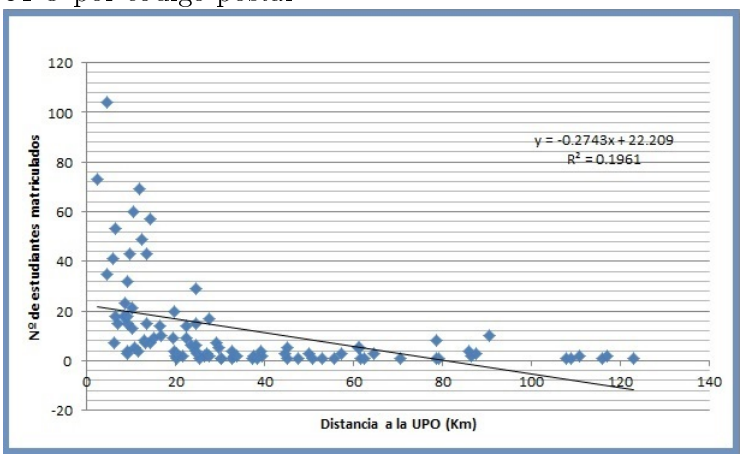


Figura 5.16).

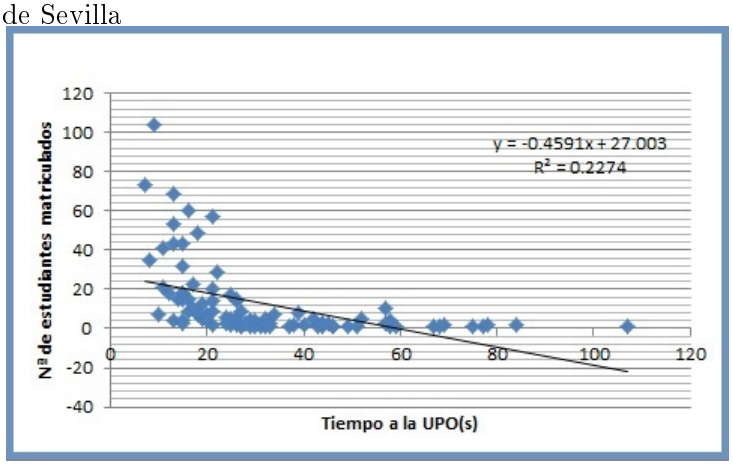
- Sevilla Capital por código postal

  En el caso de los códigos postales, esta correlación aumenta un poco con respecto al caso anterior (el del estudio por distritos), pero sigue siendo muy débil (ver la Figura 5.17).

- Conclusión

  En este apartado llegamos a obtener las mismas conclusiones que en el anterior (relativo a las distancias), si bien siendo un poco mayores las correlaciones con la variable "tiempo" (y el número de estudiantes matriculados). Como ambas variables (distancia y tiempo) están muy fuertemente correlacionadas, podría considerarse la utilización exclusiva de una de las dos (preferiblemente el tiempo) en los modelos que traten de explicar la mayor aparición de estudiantes en un municipio o barrio. No obstante, parece lógico pensar que la distancia en tiempo no siempre debe medirse mediante el uso de automóviles privados (sobre todo, tratándose de jóvenes, algunos de los cuales no tienen acceso a ese tipo de transporte y utilizan, en su lugar,

Figura 5.13: Relación entre el tiempo a la Universidad y los estudiantes de la FCE de la UPO de Sevilla



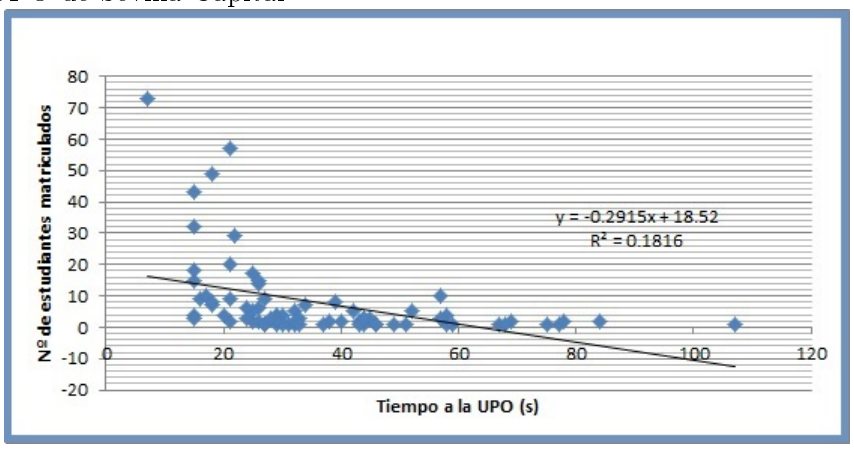
bicicletas o servicios públicos como autobús y metro.

A continuación, trataremos de incorporar la posibilidad de uso de transporte público cómodo como variable que explique parcialmente la elección de universidad por parte de los futuros alumnos.

Influencia de la línea de metro en servicio

Como podemos observar en los resultados presentados en la Tabla 5.17, el mayor número de estudiantes provienen a la UPO de zonas cuyos códigos postales corresponden a distritos donde existe alguna parada del metro de Sevilla. En las Figuras 5.19 (mapa de Sevilla por códigos postales) y 5.19 (proporción de estudiantes en la FCE de la UPO en cada zona) se aprecia la proporción de estudiantes que vienen a la UPO. La conclusión lógica es que el metro influye positivamente en la captación de alumnos; de hecho, parece probado que realiza una labor importante como transporte para los estudiantes.

Figura 5.14: Relación entre el tiempo a la Universidad y los estudiantes de la FCE de la UPO de Sevilla Capital



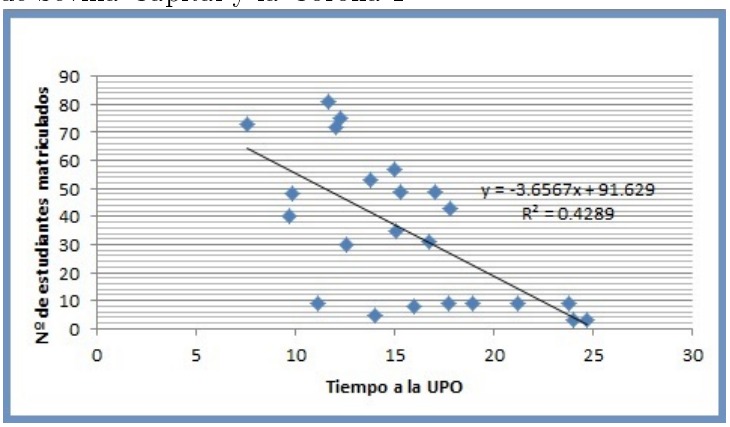
Posibles mejoras para futuros análisis relacionados

A continuación trataremos de proporcionar, mediante el uso de RNA, algunas claves que permitan encontrar los ámbitos (grupos de municipios o distritos) más apropiados para análisis similares a los presentados anteriormente. Además, se propondrá el desarrollo de aplicaciones que pudieran resultar de interés para el ámbito universitario.

La primera idea que sugerimos aquí sería la utilización de una RNA para clasificar a los alumnos según 2 o 3 de las variables anteriores (población, kilómetros y tiempo). También se podrían clasificar los municipios según 2, 3 o 4 de las variables (población, kilómetros, tiempo y alumnos en la FCE de la UPO). Las clasificaciones anteriores podrían permitir una repetición de los análisis previos (concretamente los referidos a la influencia del metro o de otras infraestructuras como el carril bici) utilizando grupos más homogéneos (y objetivamente determinados) que los anteriores.

Finalmente, aprovechamos estas líneas para sugerir a las Universidades el diseño de una herramienta informática que intente predecir la Universidad

Figura 5.15: Relación entre el tiempo a la Universidad y los estudiantes de la FCE de la UPO de Sevilla Capital y la Corona 1



que elegirá cada estudiante en función de su procedencia o, al menos, que intente estimar el porcentaje de alumnos que debería proceder de cada municipio o distrito; creemos que la información generada podría ser útil tanto para una Universidad ya implantada como para una de nueva creación.

Figura 5.16: Relación entre el tiempo a la Universidad y los estudiantes de la FCE de la UPO de Sevilla por distritos

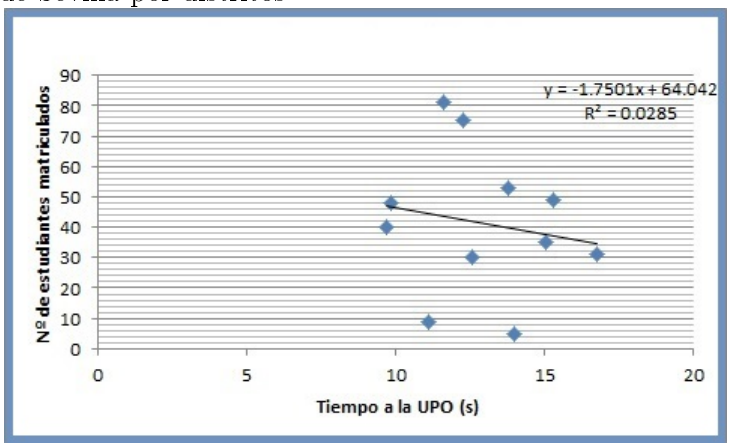


Figura 5.17: Relación entre el tiempo a la Universidad y los estudiantes de la FCE de la UPO de Sevilla por código postal

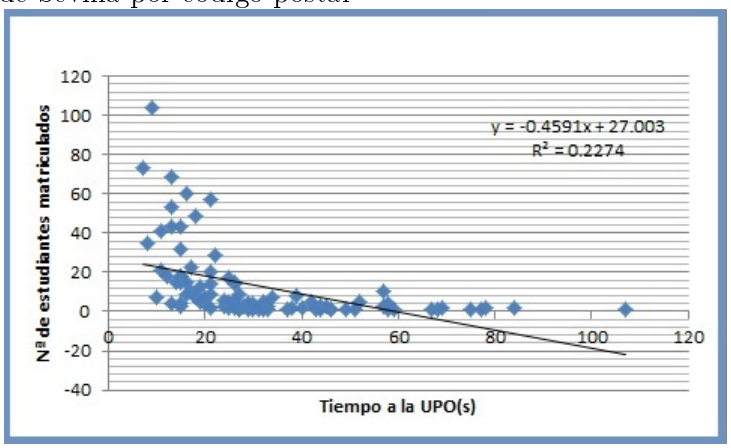


Tabla 5.17: Distribución porcentual del nº de estudiantes según la existencia de metro y por código postal

C.P.	Metro	% EST.
41001	No	2,30 %
41002	No	0,88 %
41003	No	4,07 %
41005	Sí	3,19 %
41006	Sí	6,19 %
41007	No	3,72 %
41008	No	3,19 %
41009	No	2,65 %
41010	No	7,61 %

C.P.	Metro	% EST.
41011	Sí	10,62 %
41012	Sí	7,26 %
41013	Sí	18,41 %
41014	No	1,24 %
41015	No	2,48 %
41016	No	1,24 %
41018	Sí	9,38 %
41019	No	0,71 %
41020	No	12,21 %

Fuente: elaboración propia

Figura 5.18: Mapa de Sevilla con las paradas del metro - línea 1

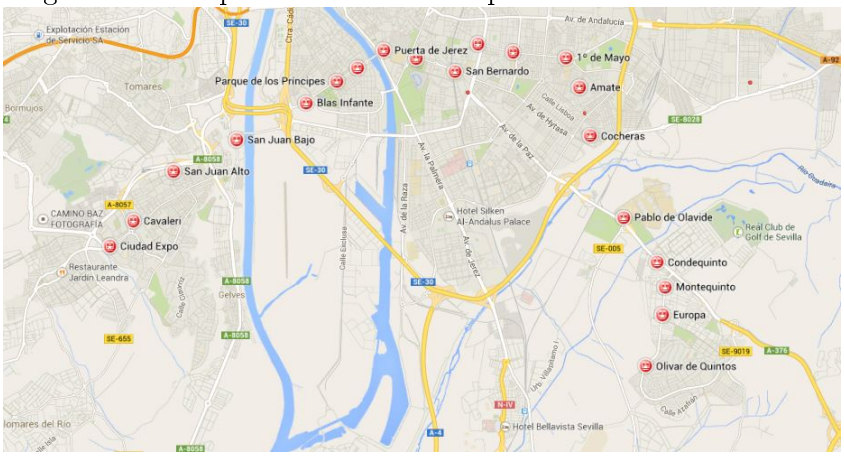


Figura 5.19: Mapa de Sevilla por códigos postales

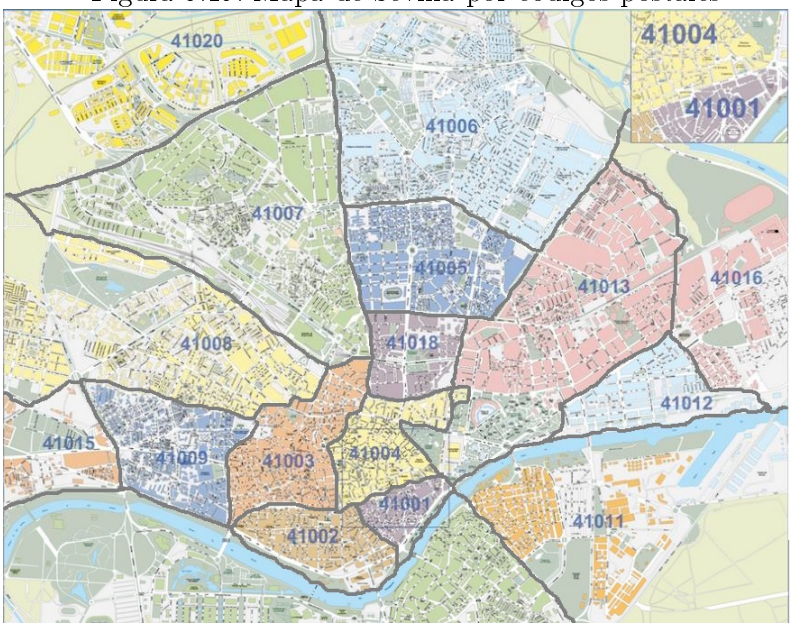
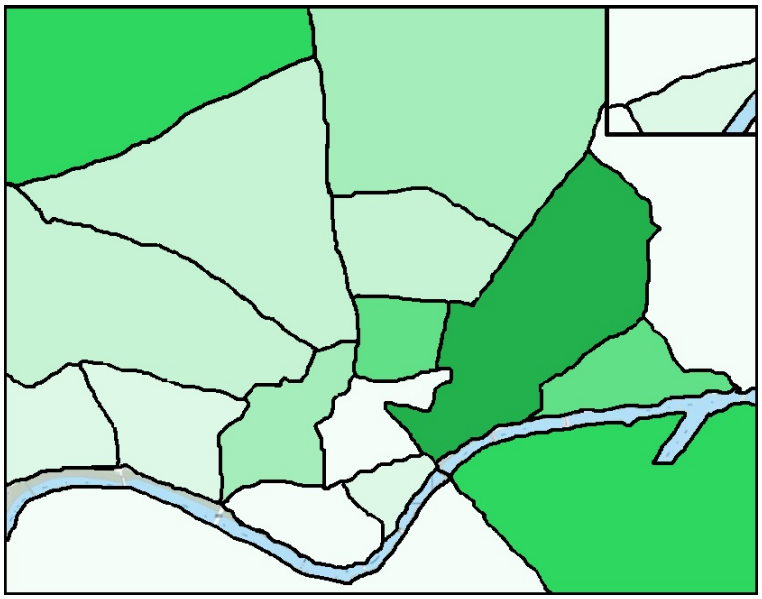


Figura 5.20: Mapa de Sevilla según el porcentaje de estudiantes de la FCE de la UPO



Capítulo 6

Aplicación

El principal objetivo de este trabajo era analizar la relación entre Educación y Economía. En concreto, uno de los aspectos más interesantes trataba de determinar qué variables afectaban al rendimiento de los estudiantes universitarios y, en especial, inferir la influencia que variables de índole económico pueden ejercer sobre el rendimiento académico. Sin embargo, la complejidad de dicha influencia y de la propia relación entre las variables explicativas hace que este objetivo no sea nada sencillo. Como ya se ha comentado, numerosos autores han tratado de explicar el éxito académico a través de diferentes variables sin lograrlo por completo. Entre otros motivos, el objeto de estudio se ha resistido por la incapacidad de las técnicas tradicionales y por la escased de datos suficientemente representativos y completos.

Atendiendo al conjunto de variables recopiladas para esta tesis, se puede apreciar que tanto la base de datos educativa como la base de datos de variables socio-económicas son densas y complejas. En un paso previo, se realizaron diferentes análisis estadísticos hasta el nivel permitido por las técnicas habituales y por la propia base de datos, llegando a concluir que el conjunto de variables está

fuertemente interrelacionado y que cada una de ellas ofrece una información no exhaustiva de todos los casos. Es decir, aparte del problema de la complejidad, del análisis descriptivo preliminar se obtuvo que los valores de muchas de las variables no se conocían para todos los individuos, existiendo mucha información perdida por diversos motivos y que, en su mayoría, no existía una solución apropiada para su recuperación.

En el siguiente apartado de este capítulo se presenta el cálculo de un indicador que nos ayudará a estimar el rendimiento académico de cada estudiante, a pesar de las dificultades ya comentadas.

6.1. Índice del rendimiento académico δ

El problema que se desea afrontar a continuación consiste en medir el rendimiento académico de cada estudiante, teniendo en cuenta aspectos objetivos y de modo que el indicador definido sirva para realizar análisis posteriores de interés. Para ello, como los individuos que se quieren comparar rara vez poseen valores para todas las variables del estudio, se utilizará la metodología aportada en la subsección 3.1.6, complementándola posteriormente con la definición de un indicador. De forma somera, para cada individuo se definirá una magnitud que establecerá el valor del nivel de rendimiento académico del estudiante hasta el momento de la medición, esté al comienzo de sus estudios, al final o en un punto intermedio.

Se considera una población finita de N individuos, todos ellos independientes. De acuerdo con la notación anterior, la información conocida del estudiante X_i viene recogida en r_i variables, con $r_i \in 0, 1, ..., n$, donde n es el número máximo de variables disponibles para el conjunto global de los estudiantes. En este apartado consideraremos que las variantes relevantes son de dos tipos, que pueden incluirse

en el conjunto de notas (en la convocatoria en que se aprobó la asignatura) o en el conjunto de convocatorias agotadas para cada asignatura (bien por haber suspendido, por haber aprobado o por no haberse presentado). Obviamente, según la notación anterior, el número de asignaturas disponibles será $\frac{n}{2}$. Para cada estudiante, habrá asignaturas de las que se tengan los dos valores (nota y convocatoria) y otras de las que solo se disponga de la convocatoria (porque el alumno no ha llegado a aprobar la asignatura en el momento del estudio: por haber suspendido o no haberse presentado). De esta aproximación puede criticarse que desecha las calificaciones de los alumnos suspensos, pero dicha elipsis es consciente, pues un alumno puede suspender una asignatura por muy diversos motivos (como no haber superado el mínimo en alguna de las partes, entre otros muchos criterios académicos posibles) y tampoco es fácil discernir si un alumno que no se presentó a examen ha rendido más o menos que otro compañero que, presentándose, suspende claramente. En nuestra opinión, las calificaciones de los alumnos suspensos no son muy representativas para realizar un análisis cuantitativo.

Así, el número de variables correspondientes a las notas, que denotaremos r_i^1 será siempre menor o igual que el de las convocatorias agotadas, r_i^2; en resumen: $0 \leq r_i^1 \leq r_i^2 \leq \frac{n}{2}$ y, consecuentemente, $0 \leq r_i = r_i^1 + r_i^2 \leq n$.

El primer problema con el que nos podemos encontrar a la hora de calcular el índice del rendimiento academico (δ) se puede pensar desde dos puntos de vista: hay una enorme diversidad de variables conocidas para cada individuo (según las asignaturas cursadas por el estudiante, así será su conjunto de variables) o faltan datos en el conjunto de individuos (si se considera que cada individuo debería tener datos de todas las variables posibles). Precisamente esta posibilidad es la que se abordó en la subsección 3.1.6.

En general, el conjunto de los N individuos se puede clasificar en p grupos, atendiendo al número de variables establecidas. Dicho número de grupos viene

dado por:

$$p = \sum_{i=1}^{n} a + n = \frac{(n+3)\,n}{2}, \text{ con } p \geq 2, \tag{6.1}$$

lo que se demuestra fácilmente por inducción:

Sea n un número par mayor o igual que 2.

1. Si $n = 2$,
$$p = \sum_{i=1}^{n} a + 2 = 1 + 2 + 2 = \frac{2\,(2+3)}{2}$$

2. Supongamos que es cierto para n:
$$p = \sum_{i=1}^{n} a + n = 1 + 2 + ... + n + n = \frac{n\,(n+3)}{2}$$

Veamos qué ocurre con $n + 1$:

$$p = \textstyle\sum_{i=1}^{n+1} a + (n+1) = 1 + 2 + ... + n + (n+1) + (n+1) = \frac{n(n+1)}{2} + n + 1 + (n+1) = \frac{n(n+1)+2(n+1)}{2} + (n+1) = \frac{(n+1)(n+2)}{2} + (n+1)$$

Una vez creados los p grupos, se establece una relación de orden total y, por tanto, una ordenación entre los diferentes grupos. Puesto que todos los elementos X_i del grupo G_k tienen los mismos valores de r_i^1 y r_i^2, podemos determinar la ordenación por la diferencia $r_i^2 - r_i^1$ o, con una notación quizá más natural, $r_k^2 - r_k^1$; es decir, la diferencia entre el número de convocatorias y el número de notas (necesariamente superiores a 5) en el grupo G_k. En caso de empate en esta diferencia, se ordena de mayor a menor según el número total de variables, $r_k^1 + r_k^2$. Este segundo criterio solo servirá para asignar un orden a los grupos, pero no afectará al índice del rendimiento académico, como sí lo hará la primera.

Una vez realizada la ordenación de los grupos, se define y calcula el coeficiente de penalización ξ. Este coeficiente penaliza a los estudiantes que tienen un mayor número de variables correspondientes a convocatorias de las que tienen notas.

Definición 6.1.1. *Según la notación anterior, r_k^2 es el número de variables de convocatorias disponibles mientras que r_k^1 es el número de variables correspondientes a notas del grupo G_k. Entonces, el coeficiente de penalización del grupo G_k es $\xi_k = 1 - \frac{2(r_k^2 - r_k^1)}{n}$, siendo $\xi \in [0,1]$.*

Una vez establecidos los p grupos G_k, se desea constituir una medida que aproxime el rendimiento académico de cada individuo. Para ello se utiliza la técnica propuesta en la subsección 3.1.6, mediante la agrupación realizada y con la ayuda de unas RNA. El principal motivo por el que se ha elegido esta técnica es que se trata de un problema de ordenación con datos faltantes, con variables relacionadas y con existencia de independiencia entre los individuos.

Una vez agrupados los N individuos en los p grupos, se diseñan y entrenan p RNA no supervisadas, una para cada grupo específico G_k. De ahí se obtienen los h_k subgrupos que se pueden crear de cada grupo G_k; para ello, se tiene en cuenta la correspondiente cota de parada definida como $\vartheta_k = \sqrt{\frac{r_k}{2}}$. (3.1.1)

Llamamos s al número total de subgrupos generados. Lógicamente, $p \leq s \leq N$. Denotamos a dichos subgrupos por S_{kh}, donde $k = 1, ..., p$ y $h = 1, ..., h_k$. Entonces, para cada subgrupo se calcula un indicador ($\overline{S_{kh}}$), que coincide con la media aritmética de las notas medias de cada individuo del subgrupo. Además de proporcionar una variable más robusta (pues no depende de una única observación), consideramos que este valor aproxima el rendimiento académico de todos los individuos del subgrupo. No obstante, es razonable tener también en cuenta el coeficiente de penalización; de ahí surge el indicador del rendimiento académico δ, que viene determinado por la siguiente fórmula:

Definición 6.1.2. *Con la notación anterior, el rendimiento académico del individuo* $X_i \in S_{kh}$ *es:*

$$\delta_i = \begin{cases} \xi_k \overline{S_{kh}} & si\ r_i^1 \neq 0 \\ \xi_k & si\ r_i^1 = 0 \end{cases}$$

donde r_{1i} *es el número de notas del estudiante* X_i *y* ξ_k *es la cota de penalización del grupo* G_k.

Obviamente, δ_i es no negativo y es menor o igual que el máximo de las medias $\overline{S_{kl}}$.

6.1.1. Cálculo del índice del rendimiento académico δ

Para determinar el rendimiento académico de cada estudiante de nuestra base de datos, se ha definido un índice δ. El valor de este índice se ha calculado atendiendo a un conjunto de variables de las cuales los distintos individuos no poseían información de todas ellas. Para establecer este indicador se ha utilizado la metodología presentada en las secciones anteriores y que se sigue a continuación.

Tenemos un conjunto de 1492 estudiantes. De cada estudiante X_i se conocen $r_i \in 0, ..., 12$ variables independientes y distribuidas de forma independiente, donde 12 es el número máximo de variables disponibles para el conjunto global de los estudiantes. Las variables están clasificadas en dos subgrupos: uno es el conjunto de notas cuando aprobó la asignatura y otro el conjunto de las convocatorias agotadas hasta aprobar la asignatura o hasta la realización de este estudio. Luego el conjunto de notas lo forman 6 variables y el conjunto de convocatorias también lo componen 6 variables, que coincide con el número de asignaturas obligatorias de Métodos Cuantitativos que cursan, como máximo, durante su estudios de Grado en la FCE de la UPO. Pero estas 12 variables, por diversos motivos, no son conocidas para todos los estudiantes, siendo posiblemente distinto el número de variables

para cada estudiante. Esta razón nos lleva a plantearnos aplicar alguna técnica para datos faltantes o perdidos; además, por la gran diversidad de variables que se tiene, se necesita una técnica más compleja a la hora de obtener un indicador útil. En este caso, para resolver el problema que nos encontramos, aplicamos unas RNA como herramientas para clasificar a los estudiantes en distintos subgrupos, de modo que cada uno de ellos tiene asignado un valor numérico como índice multiplicado por un coeficiente de personalización. La clasificación obtenida es la que se expresa a continuación y δ es el valor correspondiente al índice asignado a cada subgrupo. Luego cada estudiante está asignado a un subgrupo y tiene su valor de índice calculado para dicho subgrupo.

Una vez obtenido el índice anterior, realizar estudios más exhaustivo puede servir para comprobar la relevancia e interés del mismo; para ello, con dicho fin, se realizó un primer estudio de la correlación existente entre la variable índice y las notas del estudiante. En la Tabla 6.2, se puede apreciar que la correlación que existe entre las notas previas y el índice académico es incluso mayor que la correlación entre la media y dichas variables que supuestamente pueden utilizarse para predecir el rendimiento académico.

6.1.2. Análisis estadísticos

Una vez obtenidos los descriptivos preliminares y calculada la variable índice, se efectúan distintos análisis estadísticos, como contrastes de hipótesis, análisis multivariantes... para establecer si existen variables con diferencias significativas entre los conjuntos de datos establecidos. Una vez comprobadas las hipótesis de verificación para poder aplicar las distintas técnicas, los análisis (con un nivel de significación del 5 %) han producido los siguientes resultados que consideramos interesantes:

Índice *vs.* Curso académico

Observando el número de alumnos por curso académico, se ha establecido que sí existe diferencia significativa entre los estudiantes matriculados en cada uno de los tres cursos estudiados.

Índice *vs.* Sexo

Sexo	Frecuencia
Hombre	816
Mujer	618
Sin información	58

Atendiendo al número de hombres y mujeres se realizó un contraste de hipótesis para establecer que sí existen diferencias significativas por sexo, en cuanto al rendimiento académico. Cuando se segregan los datos por curso académico, se pueden apreciar diferencias más importantes según el sexo.

Índice *vs.* Titulación

Atendiendo a la cuatro titulaciones que se consideran en este trabajo, se observaron que existen diferencias significativas entre las distintas titulaciones. Además, realizando comparaciones múltiples, se obtuvo que existen diferencias significativas entre las cuatro titulaciones y dos a dos. Segregando los datos por curso académico y realizando el mismo análisis entre las distintas titulaciones, se obtiene la misma conclusión.

Índice *vs.* **Línea**

Al existir diferencias significativas por titulación, se decidió realizar un análisis más exhaustivo, considerando las líneas de cada titulación. Se dedujo que existen diferencias por líneas en algunas titulaciones pero en otras no. Así, en el caso particular del Grado en Finanzas y Contabilidad no se puede demostrar la existencia de diferencias significativas entre las distintas líneas de este grado.

Índice *vs.* **Grupo**

Consideramos que sería muy interesante hacer contrastes por grupos de clase, para poder valorar la influencia de unos compañeros sobre los demás, pero estos grupos son en ocasiones demasiado reducidos, por lo que no obtendríamos unos resultados suficientemente fiables sobre la existencia de diferencias significativas entre ellos. Creemos que lo más interesante de estos análisis sería ver en qué líneas en particular sí existen diferencias significativas entre los distintos grupos, para tratar de averiguar las causas de este fenómeno.

Índice *vs.* **Tipo de centro educativo de procedencia**

Una vez llevados a cabo algunos de los análisis más naturales sobre la información del estudiante en la Universidad, también puede ser interesante conocer el grado de importancia de algunas otras variables que se refieren al rendimiento del estudiante antes de entrar en la Universidad. En este caso, se estudió si existen diferencias significativas por tipo de centro donde cursaban Bachillerato, llegando al resultado de que sí existen diferencias significativas entre los estudiantes según si el centro de que provienen es público o privado concertado.

Este mismo análisis se realizó atendiendo al año académico de ingreso en la

UPO, llegando a concluir que estas diferencias se observan también en cada año académico.

Índice *vs.* Tipo de acceso

Creemos que ste resultado merece ser resaltado pues, dependiendo del curso académico en que el estudiante ingresó en la UPO, el tipo de acceso a la Universidad afecta de forma distinta al rendimiento de los estudiantes.

Algunas conclusiones

- Observando la tabla de correlaciones, se puede apreciar una correlación más fuerte entre las variables de carácter académico del estudiante y la nueva variable creada (índice δ) que con la media del estudiante en el momento del análisis. Esto nos hace pensar que la variable δ incorpora una información más amplia que la media de calificaciones, pues las variables que tradicionalmente se han utilizado para predecir el éxito del estudiante explica mejor a δ que a la mera media aritmética.

- Dependiendo del año académico en que se ingresa en la UPO, los estudiantes obtienen un rendimiento académico distinto. Es decir, parece haber factores que afectan al rendimiento del estudiante más allá de sus propias características personales. Algunos de estos factores pueden explicarse por cambios de criterios metodológicos o evaluadores (por parte de los equipos docentes) o, incluso, por cambios sociales que afectan a generaciones completas.

- Dependiendo de los estudios universitarios que se cursan, los estudiantes tienen un rendimiento diferente. Sin embargo, es probable que la principal causa de estas diferencias sea el que los estudiantes acceden con un nivel

académico mínimo distinto en cada una de las titulaciones (nota de acceso que depende de la titulación).

- Dependiendo del tipo de centro donde se estudia Bachillerato, los estudiantes obtienen un índice del rendimiento académico distinto. Esto parece indicar que hay centros donde se prepara mejor a los estudiantes, pero diferentes autores postulan que esta mejor preparación es debida a las propias características de los alumnos que estudian en uno u otro tipos de centros. Es decir, que hay centros cuyo alumnado presenta mayor propensión al aprendizaje. También parece haber diferencias en el rendimiento según el centro de procedencia (y no solo según el tipo de centro), pero los tamaños muestrales en este caso nos parecen insuficientes para extraer conclusiones serias.

- Se observa que los estudiantes pueden tener rendimientos distintos según el tipo de acceso a la Universidad, pero este suceso no ocurre de igual modo en todos los años académicos.

6.1.3. Diseño de la ruta académica óptima

Una vez concluidos los distintos análisis estadísticos preliminares, el siguiente paso que nos planteamos sería intentar establecer una "ruta académica" para cada estudiante, de modo que al entrar en la UPO se le pudiera sugerir qué titulación, qué línea... le debería resultar más recomendable. Para ello, pretendemos predecir a partir de la información previa del estudiante qué rendimiento puede obtener un alumno de nuevo ingreso en una determinada carrera universitaria.

Una vez determinadas (y elegidas) qué variables son las más apropiadas para determinar el rendimiento de un estudiante en la FCE de la UPO, se decide realizar distintos análisis (multivariantes, de varianza, de modelos logísticos, etc.) para intentar establecer una relación entre nuestro índice δ y las variables predictivas

relacionada con dicho índice. Desafortunadamente, se obtuvo que muchos de estos análisis no se podían realizar por no verificarse las condiciones previas y en otros casos ni siquiera se podían plantear (por ejemplo, por existir variables de carácter no continuo).

El conjunto de variables que vamos a considerar a la hora de analizar el fenómeno que nos ocupa es el siguiente:

1. Sexo
2. Edad
3. Distancia geográfica a la UPO desde su domicilio
4. Tiempo de llegada a la UPO desde su domicilio
5. Tipo de acceso
6. Centro de estudios previos
7. Nota del expediente académico de Bachillerato
8. Nota de la fase general del estudiante en las pruebas de acceso
9. Nota de Selectividad
10. Nota (definitiva) de acceso a la UPO
11. Población del municipio del estudiante
12. Edad media de la población del municipio del estudiante
13. Población extranjera del municipio del estudiante
14. Extensión del municipio del estudiante
15. Altitud del municipio del estudiante (respecto al nivel del mar)

16. Renta del municipio del estudiante
17. Valor catastral medio correspondiente al municipio del estudiante
18. Número de establecimientos de índole económica en el municipio del estudiante
19. Gasto energético del municipio del estudiante

En total, el conjunto de variables predictivas están compuesto por 19 variables, pero algunas de ellas las descompondremos en varias para permitir un tratamiento más adecuado, según se verá. Obviamente, la variable a explicar es nuestro indicador δ.

Volviendo al conjunto de variables explicativas estudiadas, nos encontramos que cada una de ellas está definida de una forma diferentes, con características distintas, con rangos de variación posiblemente distintos, acotadas de forma diferente y, por si fuera poco, de modo que el número de individuos observados puede variar (es decir, los valores de cada variable se conocen para un número distinto de estudiantes en cada caso).

Si atendemos al número de casos de los que conocemos todas las variables, la muestra se reduciría notablemente; para paliar este problema, se realiza una codificación específica de las variables estudiadas, obteniéndose en un conjunto más amplio de ellas. A continuación se detalla el procedimiento seguido para establecer la codificación de cada variable.

1. Sexo

   Esta variable se ha convertido en 2 variables binarias, donde una de ellas se pregunta si es mujer y la otra si es hombre; en el caso de no tener la información, el valor correspondiente sería 0 para las dos variables dicotómicas originadas.

2. Edad

   La información procedente de esta variable se ha distribuido en 9 nuevas variables, siendo todas ellas binarias. El criterio de elección de las nuevas variables ha sido el siguiente: el primer grupo lo componen los estudiantes que entran en la UPO con la edad más frecuente (17 o 18 años); el segundo grupo son estudiantes con un año más y así sucesivamente; así se han creado hasta 7 variables, siendo la última el conjunto de estudiantes que cumplen como máximo 24 años en el año natural; la octava variable corresponde a los estudiantes que tienen más de 25 años pero menos de 45 años a la hora de entrar; y la novena y última variable se refiere a los estudiantes que entran con más de 45 años (para los que hay una prueba de acceso específica).

3. Distancia geográfica a la UPO desde su domicilio

   Esta variable se ha distribuido en 7 variables, siendo estas variables todas binarias. La regla que hemos utilizado para separar los distintos subgrupos ha venido determinada, en primer lugar, por una normalización de la variable, utilizando la fórmula de normalización de la OCDE, es decir,

$$\frac{a_{ij} - \text{mín}(a_{ij})}{\text{máx}(a_{ij}) - \text{míin}(a_{ij})} \tag{6.2}$$

   para la variable j-ésima y el individuo i-ésimo. Una vez normalizada la variable, se han ordenado los valores de menor a mayor y se ha calculado la distancia entre los datos; cuando dicha distancia es mayor que $0,2$, se realiza un ajuste de grupo, es decir, se crean nuevos subgrupos grupos si existe una diferencia de al menos $0,2$ entre dos valores consecutivos. Siguiendo este procedimiento se crean, a partir de la variable distancia, un conjunto de 7 nuevas variables binarias.

4. Tiempo de llegada a la UPO desde su domicilio

   Esta variable se ha distribuido en 7 variables, siendo estas variables todas

binarias. La ley que hemos utilizado como separación de los distintos subgrupos ha venido determinada, en primer lugar, por una normalización de la variable, utilizando la fórmula 6.2; una vez normalizada la variable, se han ordenado los valores de menor a mayor y se han calculado las distancias entre los datos consecutivos; cuando la distancia es mayor que $0,2$, se realiza un ajuste de grupo, es decir, se crean grupos si existe entre dos valores consecutivos una diferencia de al menos $0,2$. Realizada esta división, se crea un conjunto de 7 variables procedentes de la variable tiempo.

5. Tipo de acceso

   Esta variable se ha convertido en 6 variables binarias, donde en cada una de ella se pregunta si se accede con una modalidad en particular, siendo el valor 1 el correspondiente a que accede de esa forma y un 0 en caso contrario. En el caso de no tener la información sobre el tipo de acceso, el valor correspondiente sería 0 para las 6 variables binarias correspondientes.

6. Centro de estudios previos

   Esta variable se ha convertido en 3 variables binarias, donde en cada una de ella se pregunta si su centro era de ese tipo en particular, siendo el valor de 1 si es de ese tipo y un 0 en caso contrario. En el caso de no tener la información, el valor correspondiente sería 0 para las 3 nuevas variables asociadas.

7. Nota del expediente académico de Bachillerato

   Esta variable se ha distribuido en 4 variables, siendo estas variables todas binarias. La ley que hemos utilizado para separar los distintos subgrupos ha venido determinada, en primer lugar, por una normalización de la variable, utilizando la fórmula 6.2; una vez normalizada la variable, se han ordenado los valores de menor a mayor y se ha dividido el conjunto de datos por cuartiles, creando 4 grupos distintos y sus correspondientes variables

dicotómicas de pertenencia.

8. Nota de la fase general del estudiante en las pruebas de acceso
   Esta variable se ha distribuido en 4 variables, siendo estas variables todas binarias. La ley que hemos utilizado para separar los distintos subgrupos ha venido determinada, en primer lugar, por una normalización de la variable, utilizando la fórmula 6.2; una vez normalizada la variable, se han ordenado los valores de menor a mayor y se han dividido por cuartiles, generando 4 grupos distintos y sus correspondientes cuatro variables dicotómicas.

9. Nota de Selectividad
   Esta variable se ha distribuido en 4 variables, siendo estas variables todas binarias. La ley que hemos utilizado para separar los distintos subgrupos ha venido determinada, en primer lugar, por una normalización de la variable, utilizando la fórmula 6.2; una vez normalizada la variable, se han ordenado los valores de menor a mayor y el conjunto de datos se ha dividido por cuartiles, creando cuatros grupos distintos y generando las correspondientes variables dicotómicas.

10. Nota (definitiva) de acceso a la UPO
    Esta variable se ha distribuido en 4 variables, siendo estas variables todas binarias. La ley que hemos utilizado para separar los distintos subgrupos ha venido determinada, en primer lugar, por una normalización de la variable, utilizando la fórmula 6.2; una vez normalizada la variable, se han ordenado los valores de menor a mayor y se ha dividido el conjunto de datos por cuartiles, creando cuatros grupos distintos y sus correspondientes variables dicotómicas.

11. Población del municipio del estudiante
    Esta variable se ha distribuido en 5 variables, siendo estas variables todas binarias. La ley que hemos utilizado para separar los distintos subgrupos ha

venido determinada, en primer lugar, por una normalización de la variable, utilizando la fórmula 6.2; una vez normalizada la variable, se han ordenado los valores de menor a mayor y se ha calculado la distancia entre los datos consecutivos; cuando la distancia es mayor que $0,2$, se realiza un ajuste de grupo, es decir, se crean grupos si existe entre dos valores consecutivos una diferencia de al menos $0,2$; realizada esta división, se ha generado para la variable población un conjunto de 5 nuevas variables.

12. Edad media de la población del municipio del estudiante
    Esta variable se ha distribuido en 4 variables, siendo estas variables todas binarias. La ley que hemos utilizado para separar los distintos subgrupos ha venido determinada, en primer lugar, por una normalización de la variable, utilizando la fórmula 6.2; una vez normalizada la variable, se han ordenado los valores de menor a mayor y se ha calculado la distancia entre los datos consecutivos; cuando dicha distancia es mayor que $0,2$, se realiza un ajuste de grupo, es decir, se crean grupos si existe entre dos valores una diferencia de al menos $0,2$; realizada esta división, se ha generado para la variable edad media de la población un conjunto de 4 nuevas variables.

13. Población extranjera del municipio del estudiante
    Esta variable se ha distribuido en 4 variables, siendo estas variables todas binarias. La ley que hemos utilizado para separar los distintos subgrupos ha venido determinada, en primer lugar, por una normalización de la variable, utilizando la fórmula 6.2; una vez normalizada la variable, se han ordenado los valores de menor a mayor y se ha calculado la distancia entre los datos; cuando la distancia es mayor que $0,2$, se realiza un ajuste de grupo, es decir, se crean grupos si existe entre dos valores consecutivos una diferencia de al menos $0,2$; realizada esta división, se ha generado para variable población extranjera un conjunto de 4 nuevas variables.

14. Extensión del municipio del estudiante Esta variable se ha distribuido en 4 variables, siendo estas variables todas binarias. La ley que hemos utilizado para separar los distintos subgrupos ha venido determinada, en primer lugar, por una normalización de la variable, utilizando la fórmula 6.2; una vez normalizada la variable, se han ordenado los valores de menor a mayor y se ha calculado la distancia entre los datos consecutivos; cuando dicha distancia es mayor que $0,2$, se realiza un ajuste de grupo, es decir, se crean grupos si existe entre dos valores consecutivos una diferencia de al menos $0,2$; realizada esta división, se ha generado en la variable extensión del municipio un conjunto de 4 nuevas variables.

15. Altitud del municipio del estudiante (respecto al nivel del mar)
    Esta variable se ha distribuido en 4 variables, siendo estas variables todas binarias. La ley que hemos utilizado para separar los distintos subgrupos ha venido determinada, en primer lugar, por una normalización de la variable, utilizando la fórmula 6.2; una vez normalizada la variable, se han ordenado los valores de menor a mayor y se ha calculado la distancia entre los datos consecutivos; cuando dicha distancia es mayor que $0,2$, se realiza un ajuste de grupo, es decir, se crean grupos si existe entre dos valores una diferencia de al menos $0,2$; realizada esta división, se ha generado en la variable altitud del municipio del estudiante un conjunto de 4 nuevas variables.

16. Renta del municipio del estudiante Esta variable se ha distribuido en 4 variables, siendo estas variables todas binarias. La ley que hemos utilizado para separar los distintos subgrupos ha venido determinada, en primer lugar, por una normalización de la variable, utilizando la fórmula 6.2; una vez normalizada la variable, se han ordenado los valores de menor a mayor y se ha calculado la distancia entre los datos consecutivos; cuando dicha distancia es mayor que $0,2$, se realiza un ajuste de grupo, es decir, se crean grupos si existe entre dos valores consecutivos una diferencia de al menos

$0,2$; realizada esta división, se ha generado en la variable renta del municipio del estudiante un conjunto de 4 nuevas variables.

17. Valor catastral medio del municipio del estudiante
    Esta variable se ha distribuido en 4 variables, siendo estas variables todas binarias. La ley que hemos utilizado para separar los distintos subgrupos ha venido determinada, en primer lugar, por una normalización de la variable, utilizando la fórmula 6.2; una vez normalizada la variable, se han ordenado los valores de menor a mayor y se ha calculado la distancia entre los datos consecutivos; cuando la distancia es mayor que $0,2$, se realiza un ajuste de grupo, es decir, se crean grupos si existe entre dos valores consecutivos una diferencia de al menos $0,2$; realizada esta división, se ha generado en la variable valor catastral del municipio un conjunto de 4 nuevas variables.

18. Número de establecimientos de índole económica en el municipio del estudiante
    Esta variable se ha distribuido en 4 variables, siendo estas variables todas binarias. La ley que hemos utilizado para separar los distintos subgrupos ha venido determinada, en primer lugar, por una normalización de la variable, utilizando la fórmula 6.2; una vez normalizada la variable, se han ordenado los valores de menor a mayor y se ha calculado la distancia entre los datos consecutivos; cuando la distancia es mayor que $0,2$, se realiza un ajuste de grupo, es decir, se crean grupos si existe entre dos valores consecutivos una diferencia de al menos $0,2$; realizada esta división, se ha generado en la variable establecimientos económicos un conjunto de 4 nuevas variables.

19. Gasto energético del municipio del estudiante Esta variable se ha distribuido en 4 variables, siendo estas variables todas binarias. La ley que hemos utilizado para sepaar los distintos subgrupos ha venido determinada, en primer lugar, por una normalización de la variable, utilizando la fórmula 6.2;

una vez normalizada la variable, se han ordenado los valores de menor a mayor y se ha calculado la distancia entre los datos consecutivos; cuando la distancia es mayor que $0,2$, se realiza un ajuste de grupo, es decir, se crean grupos si existe entre dos valores consecutivos una diferencia de al menos $0,2$; realizada esta división se ha generado en la variable gasto energético un conjunto de 4 nuevas variables.

Una vez realizadas las transformaciones oportunas, obtenemos un conjunto de 87 variables, todas ellas binarias y, lo que creemos más importante, todas ellas aportando información sobre el conjunto completo de los 1942 estudiantes.

Una vez se ha concluido la preparación de las variables predictoras, conviene considerar el cambio de magnitud en la variable índice δ, nuestra variable dependiente. Atendiendo a la definición de δ, que está acotado entre 0 y 1, su información la componen 56 valores distintos. A fin de simplificar el proceso de estimación y poder concluir con un análisis de clasificación asequible a partir de nuestro índice, se ha realizado una nueva agrupación de los valores; para ello, hemos tomado el mismo criterio que para las variables anteriores, es decir, hemos ordenados los distintos valores y hemos calculado las distancias entre los subgrupos consecutivos, agrupando de modo que la distancia entre valores del índice en un mismo subgrupo sea como máximo de $0,2$; con esto, el número de subgrupos se reduce a 10.

A continuación se propone una clasificación a partir de las variables independientes para tratar de estimar nuestra variable dependiente δ. Por las características de nuestros datos y por la gran información utilizada, no se han podido aplicar con éxito otras técnicas tradicionales, sino que se han utilizado RNA. En este caso, en particular, se ha utilizado una RNA con la siguiente estructura topológica subyacente: se trata de una red con una sola capa oculta, compuesta por 25 neuronas y cuya función de activación de la función neuronal es la sigmoide.

El resultado, una vez entrenada y probada la RNA, es que la clasificación (a partir de las 87 variables procedentes de las 19 seleccionadas) se realiza correctamente en el 77 % de los casos. Creemos que conviene destacar que con este tipo de RNA se ha podido estimar correctamente el rendimiento académico en un 77 % de los casos, a partir de variables que son conocidas antes de que el estudiante se matricule en la Universidad. Según esto, una vez obtenida la red, se pueden realizar pronósticos de éxito. Incluso, puede ser interesante estimar el porcentaje de valores incorrectos para cada categoría, es decir, para cada uno de los grupos considerados.

Atendiendo a los valores obtenidos en la Tabla 6.3, se puede apreciar que existe conjunto de grupos en los que el pronóstico se equivoca con un error inferior al 15,5 %; es decir, que el 84,4 % de los casos se clasificarían correctamente en los grupos del 1 al 9, excepto el grupo 10, en el cual el porcentaje de error está próximo al 30 %. Explicamos este suceso porque, por la distribución de nuestros datos, el conjunto de individuos con índices pertenecientes al grupo 10 (es decir, el que tiene índices cercanos a 0) tiene características diferentes al resto. Es decir, creemos que hay algunas variables fuera del modelo que precisamente explican el comportamiento de los alumnos que van a tener un índice muy deficiente. Hay estudiantes muy diversos que obtienen un rendimiento muy distinto al de sus compañeros (entendiendo aquí por compañeros a aquellos que tienen características muy similares) y este hecho es el que causa que aumente el error, porque el modelo no puede discriminar correctamente entre unos alumnos con un rendimiento próximo al esperado y otros con un rendimiento muy deficiente. También se debe comentar que esta propia naturaleza de los datos (individuos y variables conocidas) hace pensar que no se pueda obtener unos resultados más fiables que los que presentamos aquí, incluso utilizando otras metodologías más complejas.

Una vez realizado el análisis para los datos en general, también se presenta un

análisis particular para cada una de las titulaciones (ver la Tabla 6.4, por haber observado previamente la existencia de diferencias significativas entre los distintos grados.

En este caso, las RNA utilizadas tienen todas ellas la misma estructura topológica y la misma función neuronal. En particular, realizamos un análisis con una RNA con dos capas ocultas, una con 19 neuronas y la otra con 4 neuronas. Para obtener un entrenamiento más eficiente a partir de estos datos, se ha utilizado la técnica propuesta en la sección 3.3: se ha partido de una matriz de pesos iniciales con valores 0 en muchas de las conexiones entre los valores de entrada y las neuronas de la primera capa; es decir, la primera neurona de la primera capa oculta refleja la información procedente de las variables con características relativas al sexo (por lo que solamente tiene conexiones procedentes de las neuronas de entrada correspondientes al sexo del individuo) y los demás pesos serían nulos. Las conexiones no nulas de la capa de entrada con la segunda neurona de la primera capa oculta serían las correspondientes a la variable edad (en este caso, todas las conexiones con variables que no correspondan con el conjunto de variables de la edad del estudiante tiene un valor de 0 en el peso inicial) y así sucesivamente. De esta forma, realizando el procedimiento de anulación de pesos para el conjunto de todas las variables, se obtiene que el conjunto de 87 variables se relaciona con 19 neuronas en la primera capa oculta. Desde dicha capa oculta a la siguiente ya se consideran todas las conexiones posibles. En concreto, los resultados obtenidos por titulación son los que se presentan en la Tabla 6.4.

Como se puede apreciar en los resultados anteriores (ver la Tabla 6.4), al separar por titulación se suele reducir el error que se comete; en nuestro ejemplo, los estudiantes de GFC se clasifican correctamente en el $79,40\,\%$ de los casos (incrementándose en un dos por ciento el éxito respecto a la clasificación general) y para el resto de titulaciones también se consigue reducir el error.

Tabla 6.1: Clasificación de los subgrupos y valor del índice

S_{kh}	δ
S_{11}	1
S_{31}	0,942596573
S_{51}	0,927615433
S_{61}	0,903822445
S_{21}	0,896541824
S_{12}	0,831082855
S_{41}	0,820071843
S_{13}	0,761618025
S_{22}	0,751840806
S_{23}	0,729388264
S_{32}	0,720403012
S_{42}	0,71121944
S_{52}	0,708612261
S_{81}	0,700904037
S_{71}	0,68706158
S_{62}	0,666694808
S_{91}	0,657327041
S_{72}	0,620216272
S_{101}	0,596216409

S_{kl}	δ
S_{102}	0,59073055
S_{82}	0,579786986
S_{111}	0,577879792
S_{73}	0,568923564
S_{92}	0,56835598
S_{112}	0,565033825
S_{141}	0,518811649
S_{131}	0,482944065
S_{132}	0,478484945
S_{161}	0,458581896
S_{151}	0,452432406
S_{162}	0,447210338
S_{142}	0,444430527
S_{143}	0,442510661
S_{133}	0,438601177
S_{163}	0,429403044
S_{152}	0,427583911
S_{181}	0,369086985
S_{201}	0,34363091

S_{kl}	δ
S_{191}	0,3435888
S_{192}	0,336859586
S_{182}	0,335309077
S_{202}	0,329609627
S_{203}	0,322073218
S_{204}	0,32167734
S_{231}	0,260742438
S_{221}	0,234129192
S_{232}	0,205494686
S_{251}	0,107126738
S_{121}	0,101062961
S_{171}	0,080850368
S_{211}	0,060637776
S_{241}	0,040425184
S_{261}	0,020212592
S_{281}	0

Tabla 6.2: Correlación entre variables predictivas, media aritmética e índice

	Expediente	Fase general	Prueba acceso	Definitiva	Media
Índice δ	0,524	0,396	0,374	0,473	0,634
Media	0,439	0,395	0,230	0,432	1

Tabla 6.3: Porcentaje de pronósticos incorrectos para cada uno de los grupos (establecidos según δ)

Grupo	% incorrectos entrenamiento	% incorrectos validación
1	2,2 %	1,4 %
2	1,6 %	1,4 %
3	3,8 %	2,8 %
4	4,8 %	3,5 %
5	11,5 %	15,4 %
6	13,1 %	15,4 %
7	15,6 %	15,4 %
8	6,4 %	8,4 %
9	1,3 %	0,7 %
10	29 %	29,4 %

Tabla 6.4: Resultado del entrenamiento de una RNA con dos capas ocultas (para estimar el rendimiento académico por titulación)

Titulación	Error	Tiempo
GADE	0,2141	2287,82
GADE-GD	0,2284	1788,55
GFC	0,2060	2077,95
GFC-GD	0,2135	2253,93

6.2. Comprobación de resultados mediante fsQCA

Después de aplicar a nuestro problema algunas mejoras metodológicas diseñadas *ad hoc*, creemos conveniente validar los resultados mediante otra técnica que lo permita. Sin embargo, ya hemos comentado las dificultades de aplicar métodos tradicionales al problema que estamos afrontando, por lo que nos conformaremos con hacer una comprobación parcial, de modo que, al menos, podamos afirmar que nuestros resultados son razonables, en cuanto a que no son incompatibles con los obtenidos siguiendo otras estrategias.

En concreto, de todas las técnicas posibles, nos ha parecido coherente buscar alguna que esté, de algún modo, inspirada en la inteligencia artificial, por aquello de que las RNA también pueden considerarse parte de dicho campo. Actualmente se reconocen al menos cinco formas de inteligencia artificial inspiradas en el funcionamiento del cerebro humano:

1. Ejecución automática de una respuesta predeterminada a cada posible entrada: este tipo sería el más básico y el análogo a los actos reflejos de los seres vivos.

2. Previsión de un conjunto de estados producidos por las acciones posibles y posterior búsqueda del estado efectivamente sucedido.

3. Algoritmos genéticos, que están más bien inspirados en el proceso de evolución de los seres vivos que se produce por la modificación y combinación progresiva de las cadenas de ADN.

4. RNA, que imitan el funcionamiento físico del cerebro de animales y humanos, sobre todo en lo que corresponde al aprendizaje y a la respuesta ante situaciones imprevistas de antemano.

5. Razonamiento mediante una lógica formal, que sería lo más análogo al pensamiento abstracto humano, pero la lógica a veces no puede aplicarse en su versión binaria, sino que debe ser "difusa" o "borrosa" (del inglés, *fuzzy*); hay que tener en cuenta que el cerebro humano es capaz de atender a la imprecisión de la realidad y en ocasiones tiene que medir características difíciles de cuantificar (qué es algo lejano, pobre, caro...).

En este último tipo de inteligencia artificial es donde creemos que mejor se engloba una técnica novedosa denominada Análisis Cualitativo Comparativo Difuso (fsQCA), que trataremos de explicar brevemente y aplicar posteriormente para validar nuestros resultados obtenidos con la ayuda de las RNA.

6.2.1. Descripción de la técnica fsQCA

En lo que sigue, para realizar una introducción superficial al fsQCA, se recurre al material presentado en [44], aunque hay varios artículos recientes que también pueden servir para hacerse una idea de la técnica (se pueden consultar, por ejemplo, [116], [149], [150] y [3]).

Creemos que hay dos aspectos clave para entender el interés de fsQCA en el ámbito de la Educación, de la Economía, de la Empresa y, en general, de las Ciencias Sociales: por una parte, que permite extraer conclusiones de los casos particulares (desde este punto de vista, se trata del equivalente cuantitativo del método del caso en Empresa), no como las técnicas estadísticas tradicionales, que no están diseñadas para justificar la validez de los estudios sobre muestras reducidas sino para operar bajo el paraguas de las propiedades asintóticas; por otro lado, facilita la incorporación de valoraciones imprecisas (variables subjetivas o de difícil medida exacta, variables en las que no está muy seguro de los valores reales que toma, etc.), obteniéndose en muchos casos relaciones no simétricas, es

decir, que pueden detectarse causas y consecuencias sin que necesariamente se estén produciendo relaciones de equivalencia (sino solo condiciones necesarias o suficientes).

Creemos que la diferencia fundamental entre la lógica tradicional (de dos valores de verdad: verdadero y falso) y la difusa se podría comparar con la diferencia entre precisión y relevancia (o significatividad); es decir, a veces no se necesita toda la información o que toda la información sea precisa o exacta... sino que solo es necesario contar con la información que tiene realmente importancia. Este aspecto lo aporta la Lógica Difusa, en la que determinadas variaciones en los datos o resultados son relevantes mientras que otras variaciones no tienen la menor importancia.

Según [149], fsQCA surge para evitar los numerosos problemas que trae consigo la aplicación indebida de técnicas tradicionales como, por ejemplo, la Regresión Lineal Múltiple. Woodside [149] sostiene que la Regresión Lineal hace que los investigadores piensen de un determinado modo, que no siempre es el más apropiado. Por otro lado, afirma que en las regresiones suele confundirse ajuste (lo que realmente se hace) con predicción (lo que se desearía hacer).

Aparte de la crítica anterior, creemos que fsQCA es una técnica interesante para analizar conjuntamente variables de diferentes tipos (aunque se requieren transformaciones) o cuando se necesita incorporar características cuantitativas continuas junto con otras discretas o cualitativas/categóricas.

Al contrario de lo que ocurre en otras técnicas, con fsQCA no es necesario suponer independencia entre las variables explicativas y tampoco supone la existencia de relaciones causa-efecto (pues se considera una lógica asimétrica). De acuerdo con [116] y [150], el "efecto neto" no siempre es un concepto útil o válido.

Es más, tampoco es necesario suponer linealidad u otros tipos de relaciones *a*

priori entre las variables explicativas y las explicadas.

Finalmente, fsQCA permite conseguir significatividad con pocas observaciones. Creemos que esto es muy importante, porque es frecuente encontrar trabajos publicados en revistas prestigiosas en los que se extraen conclusiones con unos valores de los estadísticos muy poco significativos (R^2 o p-valor excesivamente bajos).

Antes de comentar la técnica, creemos pertinente realizar algunos comentarios sobre su origen. Por eso, a continuación trataremos de resumir brevemente los principales fundamentos del Análisis Cualitativo Comparativo (QCA).

Fue desarrollado por el prestigioso sociólogo Charles C. Ragin. Primero, antes de 1990, desarrolló la técnica denominada csQCA (de las iniciales de *crisp set*, relativa a la Lógica Booleana). Se trataba de "repensar" tipos de problemas. No se basaba en la idea de correlación sino en buscar relaciones lógicas entre "condiciones causales". Es decir, se trata a los casos como "configuraciones de causas" y se valora cuáles de dichas configuraciones tienen una influencia en los resultados que se desea analizar.

En cuanto al fsQCA, lógicamente, tiene unas hipótesis de aplicación (aunque parece bastante razonables en el caso que nos ocupa):

- Las consecuencias no suelen serlo de una sola causa, sino de una combinación de ellas.
- Diferentes combinaciones de causas pueden proporcionar el mismo resultado final.
- Normalmente, no es posible tener casos de todas y cada una de las combinaciones posibles de causas, pero eso no debe ser impedimento para extraer conclusiones lógicas.

- Las relaciones causales pueden no ser simétricas; es decir, algo puede ser causa sin ser la única y algo puede ser consecuencia sin ser la única. Desde otro punto de vista, una combinación causal no suele ser suficiente al 100 %.

- Un mismo conjunto de casos no debería utilizarse para explicar diferentes objetivos o diferentes resultados.

Hay determinadas circunstancias que parecen recomendar la aplicación de fsQCA. Comenzamos comentando que fsQCA está a medio camino entre lo cualitativo y lo cuantitativo. Es una técnica que proporciona relaciones de causalidad difusas entre determinadas configuraciones y ciertos resultados. Presta más atención a los casos que a las variables (las variables relevantes se cambian por casos (o *paths*) relevantes. Por todo lo anterior, fsQCA es una técnica apropiada y particularmente potente cuando se estudian sistemas grandes y complejos, donde la interferencia de ocurrencias y de variables es importante. Algunas áreas (no excluyentes) de aplicaciones destacadas y actuales de fsQCA incluyen: descubrir patrones ocultos en los datos cualitativos, habitualmente mediante el uso de ordenadores; estudios de Sociología, donde las variables son usualmente subjetivas; análisis de datos para la toma de decisiones (empresariales...); etc.

En cuanto al sentido de la incorporación de la Lógica Difusa al QCA (o csQCA), creemos que es justo reconocer que fsQCA también parte parte de la Lógica Booleana, pero la mejora o potencia. Así, se asigna a cada individuo un "índice/grado de pertenencia" al grupo (de modo que lo cualitativo se hace cuantitativo) que verifica las condiciones (en inglés, *recipe*, que podemos traducir por "receta" o "fórmula"). Un error muy común es considerar que este grado de pertenencia es una probabilidad. No se trata de eso, sino de asumir que cada individuo puede participar parcialmente de las características de un grupo (definido por la correspondiente receta).

Tras la aplicación de fsQCA, no siempre es posible encontrar equivalencias en los datos, pero a menudo sí es posible determinar condiciones necesarias o suficientes (o que lo son casi siempre).

En cierto modo, fsQCA se opone a la teoría de indicadores (sobre todo los unidimensionales), pues un indicador es un "output" para ordenar mientras que fsQCA no trata de dar puntuaciones a los individuos en la salida sino en la entrada.

Proceso de aplicación de fsQCA

El paso previo lógico consiste en determinar el problema y elegir los datos apropiados. Ha de tenerse en cuenta que el conjunto de casos puede variar a lo largo del proceso (normalmente se reduce). Después, conviene comprobar si los datos son "razonables"; hay que eliminar las partes de los datos que puedan ser problemáticas; a veces es pertinente dividir el conjunto de datos; debe comprobarse si el número de datos es adecuado (no muy grande ni muy pequeño...); etc.

A continuación, convertiríamos las k variables/características en "difusas". Para dar este paso, que luego explicaremos bajo el nombre de "calibrado", suele ser necesario determinar el grado de pertenencia de cada caso a cada clase. Tanto este paso como los siguientes, pueden realizarse con la ayuda de un programa específico que puede encontrarse en `http://fsqca.com`.

Con el paso anterior, es posible establecer la "truth table" o tabla de configuraciones (sin configuraciones contradictorias), con k términos (cada uno o su complementario, conectados todos por "y lógicos").

Una vez establecida la tabla de configuraciones, hay que evaluar las 2^k configuraciones (son del tipo "si se da la configuración, entonces se obtiene el resultado"),

junto con sus complementarios, estableciendo pertenencias.

No todas las configuraciones son significativas. Deben utilizarse el número de casos (eliminando los de poca frecuencia) y la Lógica Booleana (para prescindir de las cláusulas redundantes) para reducir las 2^k configuraciones (en números absolutos y en términos que tenga cada condición, si es posible).

Finalmente, el investigador debe seleccionar las reglas con adecuadas "coherencia" y "cobertura" para extraer las conclusiones pertinentes e interpretarlas. Enseguida comentaremos qué coherencias y coberturas pueden considerarse razonables en el tipo de estudios que nos ocupan.

Algunos elementos concretos relevantes del fsQCA

Aunque no deseamos extendernos en exceso en la explicación de esta técnica, que utilizamos simplemente para validación, a continuación t rataremos de realizar algunas explicaciones sobre el proceso anterior, a fin de poder facilitar su comprensión para los lectores que aún no estén familiarizados con ella.

Comenzaremos por el calibrado (o ajuste). Es como una normalización de los datos (tanto en las entradas como en las salidas), que se puede hacer en forma "binaria", "de intervalo" o "fuzzy". Como ya se ha comentado, consiste en estimar el grado de pertenencia de cada caso al grupo (o *recipe*). En el caso "fuzzy", que es el que más nos interesa, este es a menudo el punto más subjetivo del análisis, pues el investigador fija dónde están el 5 %, el 50 % y el 95 % de la distribución de pertenencia (y, en ocasiones, puede que no exista un criterio claro y objetivo para fijar dichos límites). Con el calibrado, las características se convierten en variables; en cierto modo, podríamos decir que lo cualitativo se hace cuantitativo, lo discreto se convierte en continuo... Pero debemos recordar que las variables en escala ordinal o de intervalo se convierten en porcentaje de pertenencia, luego las

variables se convierten de alguna manera en algo categórico; es decir, que también lo cuantitativo se hace aquí cualitativo.

En el fondo, lo que se hace con la calibración es distinguir entre variación relevante y variación irrelevante (lo que, en opinión de los defensores de fsQCA, hace que el "punto subjetivo" tenga menos importancia que en las regresiones o en los indicadores). Esto enlaza con el comentario que hacíamos antes sobre precisión y relevancia.

Creemos interesante comentar que en un grupo con varias características consideradas, la pertenencia de un individuo a dicho grupo coincide con el menor valor de los de cada una de las variables individuales (sin necesidad de minorarlo).

Veamos ahora algo sobre la coherencia (o *consistency*). Responde hasta qué punto (o grado) es coherente la hipótesis o el enunciado. Dicho con otras palabras, explica hasta qué grado los casos comparten características del grupo de salida; es decir, el grado en que la pertenencia en la solución es subconjunto de la salida. Se le puede encontrar cierto parecido con la correlación, pero solo en un sentido. Se suele exigir que sea mayor de $0,74$ para extraer conclusiones válidas.

Finalmente, explicaremos en qué consiste la cobertura (o *coverage*). Explica hasta qué punto cuenta el pertenecer a un grupo (*recipe*) de entrada para la variable dependiente (pertenecer a un grupo de salida). Tiene algo de similitud con el coeficiente de determinación R^2, pero lo que muestra en realidad es cuántos casos sustentan el resultado (esto es, el porcentaje de casos que cubre la solución). Se suele exigir que sea mayor que $0,25$ y menor que $0,65$ (menos no sería suficientemente significativo y más recomendaría el uso de algún tipo de regresión).

Una vez aplicado el análisis fsQCA, se pueden obtener 3 tipos de soluciones. Las más sencillas, simples o simplificadas son las que se suelen denominar "parsimonious"; las segundas son las "intermediate"; y las últimas se llaman "complex".

Las últimas son las que contienen más condiciones simples en una misma solución.

6.2.2. Resultados de la aplicación de fsQCA

Tras utilizar el programa fsQCA.exe sobre nuestros datos, se obtienen los siguientes resultados parciales y finales:

Primero recodificamos las variables, obteniendo las definiciones de pertenencia difusa presentadas en la Tabla 6.5.

Posteriormente obtenemos las siguientes soluciones (complejas, donde la conjunción lógica se denota por un asterisco y la tilde sirve para señalar la negación de una propiedad).

1. Primera solución

   ```
   ~km*~sg*~edad*bach*selec_g*selectiv*nota_acc*poblac*extranj*
   ~dist_cap*~altitud*renta*catastral*oficinas*energ
   ```

   Esta solución viene refrendada por los siguientes parámetros:

   a) Raw coverage = 0,321658

   b) Unique Coverage = 0,033912

   c) Consistency = 0,917068

   Observando los resultados obtenidos, podemos destacar que un estudiante que vive cerca de la UPO, tanto en distancia como en tiempo, con una edad joven, con buenas notas tanto en su expediente de Bachillerato y en las pruebas de selectividad (tanto en la parte genérica como en la especifica, luego, en consecuencia, con buena nota de acceso a la Universidad) va a tener un buen rendimiento académico en la FCE de la UPO (si cumple otras

Tabla 6.5: Variables recodificadas para fsQCA

Variable	Definición	Sup.	Med.	Inf.
Sexo	mujer	1.95	1.5	1.05
km	lejos (en distancia)	0.25	0.03	0.01
sg	lejos (en tiempo)	0.3	0.1	0.01
Edad	mayor	30	19	16
Centro	privado	1.95	1.5	1.05
Bach	buen expediente	1	0.5	0.2
Selec_Gen	buen expediente	1	0.6	0.4
Selectiv	buen expediente	0.6	0.2	0
Nota_Acc	buen expediente	1	0.4	0
Poblac	gran municipio (pobl.)	1	0.05	0
Edad_P	municipio envejecido	0.9	0.8	0.65
Extranj	municipio internacional	1	0.05	0
Extens	municipio extenso	1	0.05	0
Dist_Cap	municipio alejado	1	0.08	0
Altitud	municipio elevado	1	0.3	0
Renta	municipio próspero	1	0.5	0.3
Catastral	municipio caro	0.25	0.05	0
Oficinas	municipio mercantil	1	0.08	0
Energ	municipio industrial	1	0.08	0
Output	alto rendimiento	1	0.5	0

condiciones adicionales). Realmente, hasta aquí no parece una información muy sorprendente, pero también debemos fijarnos por las siguientes variables socioeconómicas detectadas como favorables para el rendimiento: en esta solución, su municipio tiene una población numerosa y un alto número de extranjeros, pero está relativamente cercana a la Capital y tanto el valor castastral como el de la renta del municipio son elevados, presentando también un gasto elevado de energía y numerosas oficinas de índole económico. La técnica garantiza que los estudiantes con estas características obtendrán un rendimiento académico elevado.

2. Segunda solución

```
~km*~sg*~edad*bach*selec_g*selectiv*~poblac*~extranj*~dist_cap*
~altitud*renta*catastral*~oficinas*~energ
```

Esta solución viene caracterizada por los siguientes parámetros:

a) Raw coverage = 0,260870

b) Unique Coverage = 0,043938

c) Consistency = 0,909926

Según esta segunda solución, un estudiante con las siguientes características obtendrá también un buen rendimiento académico al finalizar sus estudios (al menos en lo concerniente a las asignaturas de índole cuantitativa): el domicilio del estudiante está cercano a la UPO tanto en tiempo como en distancia, el estudiante es joven y tiene un buen rendimiento en Bachillerato y en Selectividad (tanto en la parte general como en la específica). En cuanto a las características socioeconómicas encontradas como relevantes, este estudiante vive en un municipio con muy poca población (tanto española como extranjera), que es cercano a la Capital, con una altitud cercana

al nivel del mar y con una renta y un valor catastral del municipio muy elevadas. El municipio gasta poca energía y posee pocas oficinas de carácter económico.

Las soluciones obtenidas no contradicen el resto de análisis considerados anteriormente y hacen pensar que son muchas las variables que afectan al rendimiento académico y que pueden servir para determinar perfiles educativos que implican éxito. En el caso de los dos conjuntos de características obtenidas, se comprueba que el éxito académico es posible tanto en municipios grandes como en pequeños, pero ambas soluciones se refieren a estudiantes con buen rendimiento académico previo y, lo que querríamos destacar aquí, en municipios con alto nivel económico (al menos, según la renta per capita y el valor catastral).

La siguiente solución más cercana a aparecer (que queda casi en el límite de las que se han eliminado por ser escasamente significativas) es la siguiente:

```
~sexo*~km*~sg*~edad*bach*selec_g*selectiv*nota_acc*poblac*extranj*
~dist_cap*~altitud*renta*catastral*energ
```

Los parámetros asociados a la solución son:

1. Raw coverage = 0,212633
2. Unique Coverage = 0,001411
3. Consistency = 0,904570

Como se puede comprobar, aquí también aparecen individuos con buen rendimiento previo y que habitan en municipios con alto poder adquisitivo (alta renta y alto valor catastral).

Capítulo 7

Conclusiones

A lo largo de la presente memoria se ha resumido el trabajo de investigación realizado durante un período de tutela doctoral. Inicialmente se pretendía ofrecer una metodología adecuada para valorar las influencias entre Economía y Educación, lo que ha llevado al estudio de una técnica relacionada con la inteligencia artificial y su posterior mejora en algunos aspectos concretos; también ha supuesto el estudio de la literatura relacionada con el tema; finalmente, ha supuesto la recopilación de datos adecuados y la aplicación continuada de las metodologías desarrolladas a diferentes subconjuntos de la base de datos considerada.

Aunque no se trataba de dar una solución al problema educativo ni a circunstancias económicas de nuestro ámbito, además de las conclusiones puramente metodológicas, se han alcanzado algunas otras relativas a las situaciones analizadas. Por eso, en las siguientes líneas se presenta un resumen de los aspectos que consideramos más destacados de entre las consecuencias que se han ido deduciendo del trabajo de investigación.

7.1. Resultados metodológicos

Comenzando con la técnica, se ha comprobado que las RNA constituyen una metodología muy potente y versátil, pero que queda bastante alejada de las posibilidades de la mayoría de los investigadores en Ciencias Sociales. Por una parte, parece que los expertos en RNA no suelen demostrar mucho interés porque sus avances (que ya de por sí requieren de un grado alto de especialización para acceder a su comprensión) sean completamente conocidos de forma pública; por otra, las RNA se desarrollan de una forma vertiginosa y es difícil encontrar manuales de iniciación convenientemente estructurados (y menos aún en español). Por eso, hemos redactado una especie de manual que pensamos que puede ser útil para los investigadores que deseen aproximarse al mundo de las RNA.

También se ha constatado que es posible utilizar variantes de las RNA actualmente en uso con el fin de obtener otras herramientas más ajustadas a las necesidades de los investigadores. Sin realizar un esfuerzo desproporcionado, se ofrecen varias mejoras metodológicas que confiamos que puedan ser aprovechadas e incluso potenciadas en el futuro. Estas mejoras se refieren, fundamentalmente, a la posibilidad de: analizar variables con distintas características (e incluso sacar un mayor partido de las variables dependientes), aprovechar toda la información en bases de datos con valores perdidos, utilizar la potencia de cálculo de los ordenadores actuales para elegir las RNA más útiles, etc.

Como conclusiones más destacables en la parte dedicada a las RNA, se ha visto que existen características relevantes a la hora de elegir la RNA más adecuada para tratar de resolver un problema; es decir, para realizar un análisis más exhaustivo y conveniente, se deben tener en cuenta las distintas partes de la definición aportada en la primera parte de esta memoria. Y hemos propuesto un programa que ayuda a elegir automáticamente la RNA más adecuada. También se ha pre-

sentado una forma de reducir el número de paramétros a estimar (reduciendo el número de conexiones en la primera capa de la RNA), utilizando las relaciones conocidas entre las variables consideradas. Finalmente, se han proporcionado soluciones para analizar un conjunto de datos sin reducir excesivamente el número de variables o casos, aunque alguna de las variables pudiera no estar definida de forma totalmente correcta o algunos de los casos presentaran información faltante.

También se ha utilizado otra técnica bastante novedosa (el fsQCA) para validar de algún modo algunos de los resultados obtenidos con las RNA. Consideramos que es una metodología interesante y que puede ser aprovechada en el futuro, incluso de forma integrada con las RNA.

7.2. Resultados de la aplicación

En cuanto a las enseñanzas más prácticas, referentes a la aplicación presentada en la segunda parte de la memoria, cada uno de los análisis realizados durante estos años (se presenten o no explícitamente en este documento) han proporcionado diferentes pistas para entender el fenómeno que se propone analizar: la relación estrecha entre Economía y Educación. A continuación trataremos de resumir algunas de las conclusiones que consideramos más destacables.

La primera conclusión que debemos destacar es que el problema es muy complejo, sobre todo, por la inmensa variabilidad en las relaciones entre las variables implicadas. También es considerable el número de variables en sí mismo, la distinta naturaleza de las características que hay que tener en cuenta, la dificultad de encontrar un conjunto de datos adecuado y fiable, etc.

Las variables referentes a género, a estudios previos, a pruebas de acceso o al nivel socioeconómico de los estudiantes son algunas de las que han sido utiliza-

das para explicar el rendimiento académico. También nosotros hemos considerado este tipo de variables y hemos demostrado que tienen una influencia probada (véanse [45], [47], [46], [48] y [50]). Autores como los de [124] y [1] afirman que estas variables que presentamos son incluso relevantes a la hora de poder diseñar mejoras significativas en la docencia, aunque nadie termina de explicitar qué modificaciones se siguen de las características analizadas.

Como es lógico, no todos los estudiantes responden con el mismo rendimiento académico a las mismas características personales. También se detecta una fuerte dependencia del tipo de asignatura; por ello, creemos importante estudiar las distintas habilidades y destrezas que se adquieren con cada signatura de ámbito cuantitativo y compararlas con los resultados del análisis, para tratar de relacionar los resultados alcanzados según el nivel real de una característica en particular.

No queremos olvidar que, en este trabajo, se ha definido un nuevo indicador del rendimiento académico y a cada estudiante de nuestra base de datos se le ha asignado un valor de dicho indicador, teniendo en cuenta los resultados académicos en la Universidad. Mediante el uso de RNA, se ha estimado un modelo que aproxima (consideramos que de forma adecuada) este indicador. Esto permitiría incluso realizar recomendaciones de índole académica a los estudiantes que comienzan, a partir de datos previos y características objetivas.

En concreto, observamos que se puede clasificar a un estudiante (según el nivel de rendimiento que va obtener) con un 80 % de acierto, aproximadamente. Consideramos que herramientas de este tipo (con los matices necesarios en cada caso) serían muy interesantes para las Universidades que deseen poder orientar a los estudiantes que ingresan por primera vez en la institución.

Otra consecuencia sobre la distribución de los datos utilizados es que consideramos necesario desarrollar herramientas informáticas más potentes que permitan detectar los análisis estadísticos más adecuados para cada conjunto de datos; es

decir, según la distribución de los datos, se realizaría un análisis preliminar automático que propondría las técnicas más pertinentes o rechazaría la aplicación de otras técnicas. Esta es una futura línea de investigación abierta, pues no todos los conjuntos de datos se podrían predecir de una forma adecuada con un mismo modelo, ni una única herramienta sería capaz de realizar el análisis previo más conveniente para proponer las técnicas que redujeran el error, el tiempo de cálculo, etc.

7.3. Otras propuestas de investigación

A lo largo de la memoria se han ido sugiriendo diferentes vías para proseguir el trabajo iniciado en esta tesis doctoral. Confiamos en tener la oportunidad de ir desarrollando cada uno de esos pequeños retos. Como líneas futuras de investigación, también se desea realizar nuevos análisis con un conjunto de datos que incorpore otras facultades y titulaciones, tratando de comparar si las rutas propuestas por la RNA son apropiadas en otros contextos académicos, para poder trasladar nuestros resultados a otros estudios.

Además, pretendemos mejorar la programación en Mathematica, para incorporar más funcionalidades al *sofware* ya desarrollado.

Finalmente, en lo que consideramos la línea de investigación más ambiciosa, trataríamos de incluir datos de la incorporación de los estudiantes al mercado laboral, para poder realizar una predicción sobre el futuro profesional de los alumnos y, en un siguiente paso, analizar la influencia de la Educación Superior en el rendimiento laboral e, indirectamente, en la Economía del país.

Bibliografía

[1] Adillón, R.; Boncompte, M.; Castañer, A.; Esteve, J.; Fort, J. M.; Jorba, L.; Ortí, F. J.; Purroy, P. Propuesta de mejora de la actuación docente a partir de las características del alumnado de primer curso de matemáticas en la Facultad de Economía y Empresa de la Universidad de Barcelona. *Rect@ 17* (2009).

[2] Afifi, A. A.; Elashoff, R. M. Missing observations in multivariate statistics I. Review of the literature. *Journal of the American Statistical Association 61* (1966), 595–604.

[3] Aguilera, J.; Fedriani, E. M.; Delgado, B. Institutional distance among country influences and environmental performance standardization in multinational enterprises. *Journal of Business Research* (2014).

[4] Alba, R.; Segundo, S. M. The return to education in Spain. *Economics of Education 14(2)* (1995).

[5] Anderson, J. A memory storage model utilizing spatial correlation functions. *Kybernetik 5* (1968), 113–119.

[6] Armeso, J. C. *Inteligencia Artificial. Una introducción filosófica.* Alianza Editorial, Madrid, (1996).

[7] Balkin, S. D.; Ord, J. K. Automatic neural network modeling for univariate time series. *International Journal of Forecasting 16* (2000), 509–515.

[8] Barceinas, F.; Alonso, J. L.; Raymond, J. L.; Roig, J. L. Los rendimientos de la educación en España. *Papeles de Economía Española.*(2000)

[9] Battiti, R. First-and second-order methods for learning: Between steepest descent and newton's method. *Neural Comutation* (1992).

[10] Becker, W. E.; Walstad, W. B. Econometric modelling in economic education research. *Boston: Kluwer Nijhoff Publishing* (1987).

[11] Behrman, J.; Pollak, R. A.; Taubman, P. Family resources, family size and access to financing for college education. *Journal of Political Economy 2* (1989), 398–419.

[12] Blas, J.; Losilla, J. M. Análisis de datos faltantes mediante redes neuronales artificiales. *Psicothema 12(3)* (2000), 503–510.

[13] Boswell, R. A. *HyperNewID and NewID.* Turing Institute, Glasgow, (1992).

[14] Bowman, N. A.; Bastedo, M. N. Getting on the front page: Organizational reputation, status signals, and the impact of us news and world report on student decisions. *Research in Higher Education 50(5)* (2009), 415–436.

[15] Brunello, G.; Checchi, D. School quality and family background. *IZA Discurssion Paper 702* (2003).

[16] Castejón, J. L. Estabilidad de diversos índices de eficacia de centros educativos. *Revista de Investigación Educativa* (1994), 45–60.

[17] Chakraborty, K.; Mehrotra, K.; Mohan, C. K.; Ranka, S. Forecasting the behavior of multivariate time series using neural networks. *Neural Networks* (1992), 961–970.

[18] Chen, X.; Simon, E. S.; Xiang, Y.; Kachman, M.; Andrews, P. C.; Wang, Y. Quantitative proteomics analysis of cell cycle-regulated Golgi disassembly and reassembly. *The Journal of Biological Chemistry 285*, 10 (2010), 7197–207.

[19] Cho, I. K.; Sargent, T. J. Neural networks for encodigo and adapting in dynamic economies. *Handbook of Computational Economics 1* (1996), 441–470.

[20] Cohen, M.; Grossberg, S. Absolute stability of global pattern formation and parallel memory storage by competitive neural networks. *IEEE Transactions on Systems, Man, Cybernetics SMC-13* (1983), 815–825.

[21] Cooper, J. C. B. Artificial neural netwroks versus multivariate statisticas: an application from economics. *Journal of Applied Statistics 26* (1999), 909–921.

[22] Coq, D. Crecimiento suburbano difuso y sin fin en el área metropolitana de Sevilla entre 1980 y 2010. Algunos elementos explicativos. *Scripta Nova. Revista Electrónica de Geografía y Ciencia XVI* (2012), 397.

[23] Corak, M.; Lipps, G.; Zhao, J. Family income and partipation in post-secondary education. *IZA Discurssion Paper 977* (2004).

[24] Cottrell, M.; Girard, B., Y, G.; Muller, C.; Rousset, P. Daily electrical power curves: Classification and forecasting using a Kohonen map. *From Natural to Artificial Neural Computation: Lecture Notes in Computer Science 930* (1995), 1107–1113.

[25] Cybenkot, G. Appoximation by superpositions of a sigmoidal function. *Mathematics of Control, Signal, and Systems 2* (1989), 203–314.

[26] Deboeck, G. J. *Trading on the Edge, Neural, Genetic, and Fuzzy Systems for Chaotic Financial Markets.* Wiley, (1994).

[27] Del Río, B M; Sanz, A. *Redes neuronales y sistemas difusos.* Alfaomega Grupo Editor, (2002).

[28] Dolado, J. J.; Morales, E. Which factors determine academic performance of undergraduate students in economics? *Documento de Trabajo. FEDEA 23* (2007).

[29] Domínguez, M.; Guerrero, F. M. *Guías docentes del Grado en Finanzas y Contabilidad - Curso académico 2009-2010.* Universidad Pablo de Olavide, 2009, Matemática Financiera.

[30] Domínguez, M.; Guerrero, F. M. *Guías docentes del Grado en Administración y Dirección de Empresas - Curso académico 2010-2011.* Universidad Pablo de Olavide, 2010, Matemática Financiera.

[31] Domínguez, M.; Guerrero, F. M. *Guías docentes del Grado en Finanzas y Contabilidad - Curso académico 2010-2011.* Universidad Pablo de Olavide, 2010, Matemática Financiera.

[32] Domínguez, M.; Guerrero, F. M. *Guías docentes del Grado en Administración y Dirección de Empresas - Curso académico 2011-2012.* Universidad Pablo de Olavide, 2011, Matemática Financiera.

[33] Domínguez, M.; Guerrero, F. M. *Guías docentes del Grado en Finanzas y Contabilidad - Curso académico 2011-2012.* Universidad Pablo de Olavide, 2011, Matemática Financiera.

[34] Fahlman, S. E. Faster-learning variations on back-propagation: An empirical study. *Proceeding, 1988 Connectionist Models Summer School, Morgan-Kufmann* (1988).

[35] Fahlman, S. E.; Lebiere, C. The cascade-correlation learning architecture. *Advances in Neural Information Processing Systems 2* (1990).

[36] Fedriani, E. M.; Melgar, Mª. C. *Guías docentes del Grado en Administración y Dirección de Empresas - Curso académico 2009-2010.* Universidad Pablo de Olavide, 2009, Matemática Empresarial I.

[37] Fedriani, E. M.; Melgar, Mª. C. *Guías docentes del Grado en Finanzas y Contabilidad - Curso académico 2009-2010.* Universidad Pablo de Olavide, 2009, Matemática Empresarial I.

[38] Fedriani, E. M.; Melgar, Mª. C. *Guías docentes del Grado en Administración y Dirección de Empresas - Curso académico 2009-2010.* Universidad Pablo de Olavide, 2009, Matemática Empresarial II.

[39] Fedriani, E. M. *Guías docentes del Grado en Administración y Dirección de Empresas - Curso académico 2010-2011.* Universidad Pablo de Olavide, 2010, Matemática Empresarial II.

[40] Fedriani, E. M. *Guías docentes del Grado en Finanzas y Contabilidad - Curso académico 2010-2011.* Universidad Pablo de Olavide, 2010, Matemática Empresarial I.

[41] Fedriani, E. M. *Las redes neuronales para la resolución de problemas de índoles económica o empresarial* Seminario del Área de Métodos Cuantitativos para la Economía y la Empresa, Universidad Pablo de Olavide, Sevilla (2010).

[42] Fedriani, E. M. *Guías docentes del Grado en Administración y Dirección de Empresas - Curso académico 2011-2012.* Universidad Pablo de Olavide, 2011, Matemática Empresarial II.

[43] Fedriani, E. M. *Guías docentes del Grado en Finanzas y Contabilidad - Curso académico 2011-2012.* Universidad Pablo de Olavide, 2011, Matemática Empresarial I.

[44] Fedriani, E. M. Introducción a un nuevo paradigma en el análisis de múltiples variables. *Seminario del Área de Métodos Cuantitativos para la Economía y la Empresa*, Universidad Pablo de Olavide, Sevilla (2014).

[45] Fedriani, E. M.; Romano, I. Causas del rendimientos académico en asignaturas cuantitativas de la Facultad de Ciencias Empresariales.*Anales de ASEPUMA, 18*, (2011), 1-15.

[46] Fedriani, E. M.; Romano, I. Hacia la homogeneización de los criterios de evaluación en asignaturas cuantitativas. *IX Foro sobre Evaluación de la Calidad de la Investigación y de la Educación Superior*, (2012), 50-54.

[47] Fedriani, E. M.; Romano, I. Diferencias en los resultados obtenidos en las pruebas realizadas por ordenador en Matemática Empresarial. *Anales de ASEPUMA, 18*, (2012), 1-22.

[48] Fedriani, E. M.; Romano, I. Análisis de una evaluación diversa en Matemática Empresarial. *Actas del XIV Congreso sobre enseñanza y aprendizaje de las matemáticas: Diversidad y Matemáticas, Málaga*, (2012), 247-256.

[49] Fedriani, E. M.; Romano, I. Increasing the Possibilities of Neural Networks to Face Economic Problems. *World Academy of Science, Engineering and Technology, 73*, (2013), 1592.

[50] Fedriani, E. M.; Romano, I. Un estudio sobre la relevancia de la asistencia a clases de Matemática Empresarial. *Anales de ASEPUMA, 22* (2014).

[51] Fedriani, F. *Hacia una cuantificación del fenómeno dialécto: Redes neuronales y debate político.* Tesis Doctoral, Universidad Pablo de Olavide, 2013.

[52] Fernández F.; González, C. On the profitability of technical trading rules based on artificial neural networks: Evidence from the Madrid stock market. *Economics Letters* (2000).

[53] Franses, P. H. B. F.; Draisma, G. Recognizing changing seasonal patterns using artificial neural networks. *Journal of Econometrics 81* (1997), 273–288.

[54] Franses, P. H. B. F.; Van Homelen, P. On forecasting exchange rates using neural networks. *Applied Financial Economics* (1998), 589–596.

[55] Fukushima, K. Neocognitron: A self-organizing neural network for a mechanism of pattern recognition unaffected by a shift in position. *Biological Cybernetics 36* (1980), 193–202.

[56] Fullana, J. *Una investigación sobre el éxito y el fracaso escolar desde la perspectiva de los factores de riego: implicaciones para la investigación y la práctica educativa.* Tesis Doctoral, Universidad de Girona, 1995.

[57] García, A.; Ramírez, J. M. *Guías docentes del Grado en Administración y Dirección de Empresas - Curso académico 2010-2011.* Universidad Pablo de Olavide, 2010, Estadística Empresarial I.

[58] García, A.; Ramírez, J. M. *Guías docentes del Grado en Finanzas y Contabilidad - Curso académico 2010-2011.* Universidad Pablo de Olavide, 2010, Estadística para Fiananzas II.

[59] García, A.; Ramírez, J. M. *Guías docentes del Grado en Administración y Dirección de Empresas - Curso académico 2011-2012.* Universidad Pablo de Olavide, 2011, Estadística Empresarial I.

[60] García, A.; Ramírez, J. M. *Guías docentes del Grado en Finanzas y Contabilidad - Curso académico 2011-2012.* Universidad Pablo de Olavide, 2011, Estadística para Fiananzas II.

[61] García, A.; Hinojosa, M. A. *Guías docentes del Grado en Finanzas y Contabilidad - Curso académico 2011-2012.* Universidad Pablo de Olavide, 2011, Estadística para Finanzas II.

[62] Grossberg, S. A prediction theory for some nonlinear functional-differential equations: I. learning of lists. *Journal of Mathematical Analysis and Aplications 21* (1968), 643–694.

[63] Grossberg, S. The theory of embedding fields with applications to Psychology and Neurophysiology. *New York: Rockefeller Institute of Medical Research* (1964).

[64] Haefke, C.; Helmenstein, C. Forecasting Austrian ipos: An aplication of lineal and neural network error-correction models. *Journal of Forecasting 15* (1996), 237–251.

[65] Hartley, H. O.; Hocking, R. R. The analysis of incomplete data. *Biometrics 27* (1971), 783–808.

[66] Hebb, D. Organization of behavior. *New York: John Wiley & Sons* (1949).

[67] Hiemstra, D. *Using statistical methods to create a bilingual dictionary.* PhD thesis, University of Twente, 1996.

[68] Hilera, J. R., Martínez, V. J. *Redes neuronales artificiales. Fundamentos, modelos y aplicaciones.* Ra-Ma, Madrid, 1995.

[69] Hinojosa, M. A.; García, A., *Guías docentes del Grado en Administración y Dirección de Empresas - Curso académico 2009-2010.* Universidad Pablo de Olavide, 2009, Estadística Empresarial I.

[70] Hinajosa, M. A. *Guías docentes del Grado en Administración y Dirección de Empresas - Curso académico 2010-2011.* Universidad Pablo de Olavide, 2010, Estadística Empresarial II.

[71] Hinajosa, M. A. *Guías docentes del Grado en Administración y Dirección de Empresas - Curso académico 2010-2011.* Universidad Pablo de Olavide, 2011, Estadística Empresarial II.

[72] Hopfield, J. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences 79* (1982), 2554–2558.

[73] Hopfield, R. G. The principales and practice of time series forecasting and business modelling using neural nets. *Neural Computing & Applications 1* (1993), 59–66.

[74] Hornik, K. Multilayer feedforward networks are universal approximators. *Neural Networks 2* (1989), 359–366.

[75] Hornik, K.; Stinchcombe, M.; White, H. Universal aproximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks 3* (1990), 551–560.

[76] Hull, J.; White, A. Incorporating volatility updating into the historial simulation method for value at risk. *Journal of Risk* (1998), 5–19.

[77] IECA. *Sistema de Información Multiterritorial de Andalucía (SIMA).* Junta de Andalucía, Consejería de Economía, Innovación, Ciencia y Empleo, 2013.

[78] INE. *Anuario Estadístico de España 2009.* Instituto Nacional de Estadística, 2009.

[79] INE. *Anuario Estadístico de España 2010.* Instituto Nacional de Estadística, 2010.

[80] INE. *Anuario Estadístico de España 2011.* Instituto Nacional de Estadística, 2011.

[81] Kaufman, A.; Gil, J. *Grafos neuronlaes para la economía y la gestión de Empresas.* Editorial Pirámide, Madrid, 1995.

[82] Kohonen, T. A class of randomly organized associative memories. *Acta Polytechnic Scandanavica* (1971).

[83] Lachtermacher, G.; Fuller, J. D. Backpropagation in time-series forecasting. *Journal of Forecasting 14* (1995), 381–393.

[84] Lauer, C. Family background, cohort and education: A French-German comparison based on a multivariate ordered probit model of educational attainment. *Labour Economics 10* (2003), 231–251.

[85] López, P. Variables asociadas a la gestión escolar como factores de calidad educativa. *Estudios Pedagógicos 1* (2010), 147–158.

[86] Maasoumi, E.; Khotanzad, A.; Abaye, A. Artificial neural networks for some macroeconomic series: A first report. *Econometric Reviews 13(1)* (1971).

[87] Marcenaro, O. D. *Resumen del Informe: La función de producción educativa para el caso de Andalucía: un análisis desde la perspectiva cuantitativa y cualitativa*, vol. 6. Instituto de Estadística y Cartografía de Andalucía, 2012.

[88] Marcenaro, O. D.; Navarro, M. L. *Condiciones de acceso y otras características del estudiante con determinantes del éxito en el primer curso universitario*, vol. 6. Actas XII Jornadas de la Asociación de Economía de la Educación, 2003.

[89] Martín, A. M. *Valoración de la pobreza mediante téncias de agregación de datos de diferente naturaleza.* Tesis Doctoral, Universidad Pablo de Olavide, 2005.

[90] Martínez, M.; Esteban, F.; Buxarrais, M. R. Escuela, profesorado y valores. *Revista de Educación Nº Estraordinario 2011* (2011), 95–113.

[91] Matthai, A. Estimation of parameters from incomplete data with aplication to design of sample surveys. *Sankhya 2* (1951), 145–152.

[92] McCulloch, W. S.; Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics 5* (1943), 115–133.

[93] Melgar, Mª. C. *Guías docentes del Grado en Administración y Dirección de Empresas - Curso académico 2010-2011.* Universidad Pablo de Olavide, 2010, Matemática Empresarial I.

[94] Melgar, Mª. C. *Guías docentes del Grado en Finanzas y Contabilidad - Curso académico 2010-2011.* Universidad Pablo de Olavide, 2010, Matemática Empresarial II.

[95] Melgar, Mª. C. *Guías docentes del Grado en Administración y Dirección de Empresas - Curso académico 2011-2012.* Universidad Pablo de Olavide, 2011, Matemática Empresarial I.

[96] Melgar, Mª. C. *Guías docentes del Grado en Finanzas y Contabilidad - Curso académico 2011-2012.* Universidad Pablo de Olavide, 2011, Matemática Empresarial II.

[97] Minsky, M. *Neural analog networks and the brain model problem.* PhD thesis, Princeton University, 1954.

[98] Minsky, M.; Papert, S. Perceptrons. *Cambridge: MIT Press* (1969).

[99] Montero, M. C. *Predicción del rendimiento académico. Estudio de las variables intervinientes en una muestra de alumnos de 8º de EGB con seguimiento en 2º de BUP.* Tesis Doctoral, Universidad Pontificia de Salamanca, 1990.

[100] Moshiri, S.; Cameron, N. Neural networks versus econometric model in forecasting inflation. *Journal of Forecasting 19* (2000), 201–217.

[101] Müller, B.; Reinhardt, J. *Neural Networks. An introduction.* Springer-Verlag, Berlín (1990)

[102] Nijkamp, P.; Wang, S. Winners and lossers in the European Monetary Union: A neural network analysis of spatial industrial shifts. *Publicaciones del Dept. of Spatial Economic del Tinbergen Institute 377* (1998), 1–18.

[103] Núñez, C. *La construcción de una red neuronal para el análisis de riesgos en las entidades financieras.* Tesis Doctoral, Univesidad de Sevilla, 1998.

[104] Ordaz, J. A. *Guías docentes del Grado en Administración y Dirección de Empresas - Curso académico 2010-2011.* Universidad Pablo de Olavide, 2010, Métodos Estadísticos y Econométricos en la empresa.

[105] Ordaz, J. A. *Guías docentes del Grado en Administración y Dirección de Empresas - Curso académico 2011-2012.* Universidad Pablo de Olavide, 2011, Métodos Estadísticos y Econométricos en la empresa.

[106] Ordaz, J. A. *Guías docentes del Grado en Finanzas y Contabilidad - Curso académico 2011-2012.* Universidad Pablo de Olavide, 2011, Métodos Estadísticos y Econométricos en Finanzas.

[107] Otero, J. M.; Trujillo, F. Predicción multivariante y multiperíodo de una serie temporal económica mediante una red neuronal. *Estadística Española 35(133)* (1993), 345–375.

[108] Paradas, C. T.; Hutchinson, G. Neural network forecasts of input-output technology. *Applied Economics 34* (2002), 1607–1615.

[109] Parker, D. B. Learning logic. *Center for Computational Research in Economics and Management Science, Technical Report* (1985), 47.

[110] Pedraja, F.; Salinas, J. La restricción de ponderaciones en el análisis envolvente de datos: una fórmula para mejorar la evaluación de la eficiencia. *Investigaciones Económicas 1* (1994), 365–380.

[111] PISA. *PISA 2006: Programa para la evaluación internacional de los alumnos. Informe español. Resultados y contexto.* Ministerio de Educación, Cultura y Deporte, 2008.

[112] PISA. *PISA 2009: Programa para la evaluación internacional de los alumnos. Informe español. Resultados y contexto.* Ministerio de Educación, Cultura y Deporte, 2011.

[113] PISA, *PISA 2012: Programa para la evaluación internacional de los alumnos. Informe español. Resultados y contexto.* Ministerio de Educación, Cultura y Deporte, 2014.

[114] Pitarque, A.; Ruiz, J. C. Encoding missing data in back-propagation neural networks. *Psicologica 17* (1996), 83–91.

[115] Plasmans, J. E. J.; Verkooijen, W. J. H.; Daniels, H. A. M. Estimating structural exchange rate models by artificial neural networks. *Applied Financial Economics 8(5)* (1998), 541–551.

[116] Ragin, C. *Redesigning social inquiry: Fuzzy sets and beyond.* Chicago University Press, 2008.

[117] Ramón y Cajal, S. *Histologie du système de l'homme et des vertébrés.* Maloine, París, 1911.

[118] Rao, J. N. K.; Shao, J. Jacknife variance estimation with survey data under hot deck imputation. *Biometrika 79* (1992), 811–822.

[119] Raymond, J. Convergencia real de las regiones españolas y capital humano. *Papeles de Economía Española 93* (2002), 109–121.

[120] Refenes, A. P. N.; Zapranis, A. D.; Francis, G. Stock performance modeling using neural networks: A comparative study with regression models. *Neural Networks 7* (1994), 375–388.

[121] Rice, P. The demand for post-compulsory education in the UK and the effects on educational maintenance allowances. *Economica 54* (1987), 465–475.

[122] Romany, R.; Santín, D. *El control de gasto público por incapacidad temporal mediante redes neuronales.* Ministerio de Hacienda y Administraciones Pública, 2002.

[123] Rosenblatt, F. Two theorems of statistical separability in the perceptron. *Mechanization of Thought Processes I* (1959).

[124] Ruá, A.; Redondo, R.; Martínez, C.; Fabra, M. E.; Martín, M. J.; Núñez, A. Factores del rendimiento académico en las asignaturas cuantitativas de Administración y Dirección de Empresas *Anales de ASEPUMA, 18*, (2010), 105.

[125] Rubin, D. B. Multiple imputation in sample surveys. In *American Statistical Association, Survey research methods section.*(1978)

[126] Rubin, D. B. Discussion on multiple imputation. *International Statistical Review* (2003), 619–625.

[127] Rumelhart, D., Hinton, G., Williams, R. Learning representations by backpropagating errors. *Nature 3* (1986), 533 – 536.

[128] Russell, S. A practical device to simulate the working of nervous discharges. *Journal of Animal Behaviour 3* (1913), 15.

[129] Sanchéz, A. M. *Tratamiento de series temporales desigualmente espaciadas: Aplicaciones en el ámbito de la economía.* Tesis Doctoral, Universidad Pablo de Olavide, 2011.

[130] Saradindu, G.; Kishore, N. K. Applicaion of artificial neural network for modelling of discharge inception voltage. *Electrical Insulation and Dielectric Phenomena* (1997), 508–511.

[131] Sharda, R.; Patil, R. B. Connectionist approach to time series prediction. *Journal of Intelligent Manufacturing 3* (1992), 317–323.

[132] Sharpe, P. K.; Solly, R. J. Dealing with missing values in neural network-based diagnostic systems. *Neural Computing and Applications 3* (1995), 73–77.

[133] Sjöberg, J. *Mathematica Neural Networks: Train and Analyze Neural Networks to Fit your Data.* Champaign, Illinois, 2005.

[134] Swanson, N. R.; White, H. A model-selection aproach to assessing the information in the term structure using linear models and artificial neural networks. *Journal of Business and Economic Statistics 13* (1995), 265–275.

[135] Swanson, N. R.; White, H. A model selection approach to real-time macroeconomic forecasting using linear models and artifical neural networks. *The Review of Economics and Statistic 79(4)* (1997), 540–550.

[136] Swingler, K. *Appying Nwural Networks. A Practical Guide.* Academic Press, 1996.

[137] Tang, Z.; Fishwick, P. A. Feedforward neural nets as models for time series forecasting. *Journal on Computing 5(4)* (1993), 374–385.

[138] Tejedor, F. J.; García, A. Causas del bajo rendimiento del estudiante universitario. propuesta de mejora en el marco del EEES. *Revista de Educación 342* (2007), 443–473.

[139] Tkacz, G. Neural network forecasting of Canadian GDP growth. *Internacional Journal of Forecasting 17(1)* (2001), 374–385.

[140] Torra, S., Monte, E. *Modelos Neuronales Aplicados en Economía.* Addlink Software Científico, Barcelona, España, 2013.

[141] Tres, V.; Ahmad, S.; Neuneier, R. *Training neural networks with deficient data.* Advances in neural information processing systems (NIPS), San Mateo: Morgan Kaufmann, 1994.

[142] Vamplew, P.; Adams, A. Missing values in a backpropagation neural net. *Proceedings of the 3rd. Australian Conference on Neural Networks (ACNN) I* (1993), 64–66.

[143] Verkooijen, W. J. Neural networks in economic modelling. *Center for Economic Research, Tilburg University* (1996).

[144] Vroomen, B.; Franses, P. H. B. F.; Van Nierop, E. Modeling consideration sets and brand choice using artificial neural networks. *European Jounal of Operational Research 154* (2004), 206–217.

[145] White, H. Learning in artificial neural networks: A statistical perspective. *Neural Computation 1* (1989), 425–464.

[146] White, M. E. *After the Greening. The Browning of Australia.* Kangaroo Press, Kenthurst, 1994.

[147] Widrow, B. Adaptive sampled-data systems - a statistical theory of adaptation. *Wescon Convention Record 4* (1959), 74–85.

[148] Winston, P. H. *Artificial Intelligence.* Cambridge, Mass.: Articial Intellingence Laboratory, M.I.T, 1976.

[149] Woodside, A. G. Moving beyond multiple regression analysis to algorithms: Calling for adoption of a paradigm shift from symmetric to asymmetric thinking in data analysis and crafting theory. *Journal of Business Research* (2014).

[150] Woodside, A. G.; Zhang, M. Cultural diversity and marketing transactions: Are market integration, large community size, and world religions necessary for fairness in ephemeral exchanges? *Psychology and Marketing* (2013).

Anexo A

Datos ejemplos

Descripción del problema a resolver:

A continuación de describen los diferentes casos, en particular se estudian 17 casos. En cada uno de ellos se han observado 80 variables distintas definida por un valor entre $\{0, 1, 2, 3, 4\}$, siendo el 0 el menor valorado y el 4 el máximo valorado.

Caso 1 :

$\{2., 3., 2., 3., 2., 2., 3., 2., 3., 3., 1., 0., 2., 1., 3., 1, 4., 4., 1., 3.,$
$4., 0., 3., 0., 4., 3., 4., 4., 3., 0., 4., 0., 2., 4., 1., 1., 3., 3., 1., 0.,$
$2., 2., 2., 2., 3., 2., 2., 2., 2., 2., 1., 0., 2., 1., 0., 1, 4., 4., 3., 2.,$
$3., 0., 0., 0., 4., 3., 4., 4., 3., 0., 3., 0., 2., 3., 2., 2., 1., 2., 1., 0.\}$

Caso 2 :

$\{3., 3., 2., 3., 2., 2., 2., 3., 2., 2., 1., 0., 1., 0., 4., 1, 3., 4., 2., 2.,$
$3., 0., 0., 0., 3., 2., 2., 3., 3., 0., 3., 0., 2., 3., 3., 3., 3., 3., 0., 0.,$
$2., 2., 2., 3., 2., 2., 1., 3., 3., 3., 1., 0., 1., 0., 0., 1, 3., 4., 3., 3.,$
$3., 1., 0., 1., 2., 2., 3., 3., 3., 0., 2., 0., 2., 3., 2., 2., 3., 4., 0., 0.\}$

Caso 3 :

$\{2., 3., 3., 3., 1., 3., 1., 1., 3., 2., 2., 2., 2., 0., 3., 1, 4., 4., 2., 1.,$
$1., 1., 0., 2., 1., 3., 1., 3., 3., 0., 3., 2., 3., 3., 0., 0., 2., 2., 1., 1.,$
$2., 3., 2., 3., 1., 3., 1., 2., 3., 3., 1., 1., 2., 0., 0., 1, 4., 4., 1., 1.,$
$1., 1., 0., 2., 1., 3., 1., 3., 3., 0., 3., 0., 3., 3., 3., 1., 2., 3., 2., 2.\}$

Caso 4 :

$\{2., 1., 3., 2., 1., 2., 1., 3., 2., 0., 1., 0., 1., 2., 0., 1, 3., 4., 3., 0.,$
$4., 3., 0., 1., 3., 2., 4., 3., 1., 0., 4., 0., 2., 4., 0., 0., 0., 1., 1., 0.,$
$0., 1., 1., 1., 3., 3., 0., 2., 1., 0., 2., 0., 2., 2., 0., 1, 2., 2., 2., 0.,$
$4., 0., 0., 0., 3., 2., 4., 3., 1., 0., 2., 0., 2., 2., 4., 3., 0., 3., 1., 0.\}$

Caso 5 :

{2., 3., 2., 1., 2., 2., 2., 2., 1., 0., 1., 0., 3., 0., 0., 1, 4., 4., 2., 3.,
3., 3., 0., 0., 2., 2., 3., 3., 2., 0., 4., 3., 2., 3., 0., 0., 0., 3., 2., 2.,
1., 1., 2., 2., 1., 3., 1., 2., 2., 0., 1., 0., 2., 0., 2., 1, 3., 3., 4., 2.,
1., 2., 0., 1., 3., 2., 4., 2., 2., 0., 3., 2., 3., 3., 2., 3., 0., 2., 1., 0.}

Caso 6

{2., 2., 4., 2., 1., 2., 1., 2., 1., 0., 0., 1., 2., 0., 0., 1, 3., 3., 3., 2.,
4., 1., 0., 1., 3., 2., 2., 3., 2., 0., 3., 2., 2., 2., 1., 1., 0., 2., 3., 3.,
1., 1., 2., 2., 3., 3., 0., 2., 2., 0., 0., 1., 1., 0., 3., 1, 3., 3., 1., 2.,
3., 2., 0., 1., 1., 3., 1., 3., 2., 0., 2., 2., 4., 3., 3., 3., 1., 4., 2., 2.}

Caso 7 :

{3., 3., 2., 3., 2., 2., 1., 2., 3., 1., 0., 0., 2., 0., 0., 1, 4., 3., 0., 2.,
3., 1., 2., 1., 4., 3., 3., 4., 2., 3., 3., 2., 2., 3., 3., 1., 1., 3., 2., 0.,
1., 1., 2., 2., 2., 3., 2., 2., 2., 1., 0., 0., 1., 0., 0., 1, 3., 2., 0., 3.,
4., 1., 2., 1., 2., 2., 2., 3., 2., 2., 2., 2., 3., 2., 3., 3., 1., 4., 1., 0.}

Caso 8 :

{2., 3., 2., 3., 2., 2., 1., 3., 3., 1., 0., 0., 0., 0., 0., 1, 3., 3., 0., 3.,
3., 2., 0., 2., 3., 3., 3., 3., 2., 0., 3., 0., 3., 4., 2., 2., 2., 2., 2., 2.,
1., 1., 2., 1., 1., 3., 0., 2., 1., 1., 0., 0., 0., 0., 0., 1, 2., 2., 2., 1.,
3., 0., 0., 0., 2., 2., 4., 2., 1., 0., 2., 0., 2., 3., 4., 4., 0., 3., 1., 1.}

Caso 9 :

{2., 2., 2., 2., 2., 3., 1., 2., 1., 1., 1., 1., 1., 0., 0., 1, 3., 2., 1., 2.,
1., 1., 0., 0., 2., 2., 3., 2., 3., 0., 3., 0., 2., 2., 3., 3., 1., 1., 0., 0.,
3., 3., 2., 3., 3., 2., 2., 2., 2., 2., 0., 1., 3., 1., 0., 1, 4., 3., 0., 3.,
1., 0., 2., 2., 2., 2., 3., 3., 2., 0., 4., 0., 2., 3., 1., 1., 3., 2., 2., 1.}

Caso 10 :

$\{1., 2., 2., 1., 2., 3., 2., 2., 1., 1., 0., 0., 2., 1., 0., 1, 2., 2., 1., 2.,$
$0., 0., 1., 1., 2., 2., 1., 2., 2., 0., 2., 1., 2., 2., 1., 1., 0., 3., 0., 0.,$
$2., 3., 2., 3., 3., 2., 3., 3., 3., 3., 0., 0., 3., 1., 1., 1, 3., 3., 0., 3.,$
$3., 0., 2., 1., 3., 3., 2., 3., 2., 0., 3., 0., 2., 3., 1., 1., 2., 2., 1., 0.\}$

Caso 11 :

$\{1., 2., 2., 2., 2., 2., 1., 2., 1., 1., 1., 0., 3., 0., 0., 1, 2., 2., 0., 2.,$
$2., 1., 3., 0., 2., 2., 2., 2., 3., 0., 2., 0., 2., 2., 2., 1., 0., 4., 1., 0.,$
$3., 3., 2., 3., 3., 3., 1., 3., 3., 3., 1., 0., 3., 0., 0., 1, 3., 3., 0., 1.,$
$1., 1., 0., 0., 3., 3., 1., 3., 3., 0., 3., 1., 2., 3., 0., 0., 0., 2., 2., 1.\}$

Caso 12 :

$\{2., 2., 1., 1., 2., 3., 2., 2., 2., 1., 0., 0., 2., 0., 0., 1, 2., 2., 1., 2.,$
$0., 0., 0., 0., 2., 2., 1., 2., 3., 1., 2., 1., 2., 2., 1., 1., 2., 3., 2., 0.,$
$2., 2., 3., 3., 3., 2., 2., 3., 3., 3., 0., 0., 3., 0., 0., 1, 3., 3., 1., 3.,$
$2., 1., 1., 1., 2., 2., 2., 3., 3., 0., 3., 0., 2., 3., 0., 0., 1., 4., 1., 0.\}$

Caso 13 :

$\{1., 1., 2., 2., 2., 3., 1., 2., 1., 1., 1., 0., 1., 0., 1., 1, 2., 2., 1., 2.,$
$1., 0., 0., 1., 2., 2., 1., 2., 3., 0., 2., 0., 2., 2., 2., 2., 1., 4., 1., 0.,$
$2., 2., 1., 3., 3., 2., 2., 3., 2., 2., 2., 0., 2., 0., 2., 1, 3., 3., 1., 2.,$
$1., 1., 0., 1., 3., 3., 1., 3., 3., 0., 3., 0., 2., 3., 1., 1., 3., 2., 2., 1.\}$

Caso 14 :

$$\{1., 3., 2., 2., 2., 2., 1., 3., 1., 1., 0., 0., 1., 0., 1., 1, 2., 2., 1., 2.,$$
$$1., 0., 1., 0., 2., 2., 2., 2., 2., 0., 2., 0., 2., 3., 2., 2., 1., 0., 1., 0.,$$
$$0., 1., 1., 1., 1., 3., 0., 3., 1., 0., 0., 0., 1., 0., 0., 1, 2., 2., 1., 1.,$$
$$0., 0., 0., 0., 0., 2., 1., 2., 1., 0., 2., 0., 2., 2., 2., 1., 0., 0., 0., 0.\}$$

Caso 15 :

$$\{1., 1., 1., 1., 2., 3., 1., 3., 1., 2., 0., 0., 1., 0., 0., 1, 2., 2., 0., 1.,$$
$$0., 0., 0., 1., 2., 2., 2., 2., 2., 0., 2., 0., 2., 2., 1., 2., 1., 0., 2., 2.,$$
$$2., 2., 2., 2., 2., 3., 0., 2., 0., 0., 0., 0., 0., 0., 0., 1, 2., 2., 0., 1.,$$
$$0., 0., 1., 0., 2., 2., 1., 2., 1., 0., 1., 0., 2., 2., 1., 1., 0., 1., 1., 0.\}$$

Caso 16 :

$$\{0., 3., 2., 2., 3., 3., 1., 2., 1., 1., 0., 0., 1., 0., 0., 1, 2., 2., 0., 1.,$$
$$2., 1., 1., 0., 2., 2., 2., 2., 1., 0., 2., 0., 2., 0., 2., 1., 1., 3., 3., 2.,$$
$$0., 1., 2., 3., 1., 3., 0., 3., 3., 3., 0., 0., 1., 0., 0., 1, 2., 2., 0., 2.,$$
$$3., 1., 2., 0., 2., 2., 2., 2., 1., 0., 2., 0., 3., 0., 1., 1., 0., 1., 1., 0.\}$$

Caso 17 :

$$\{2., 2., 2., 3., 1., 3., 1., 3., 1., 3., 0., 0., 1., 0., 0., 1, 2., 2., 0., 2.,$$
$$1., 1., 1., 0., 2., 2., 2., 2., 2., 0., 2., 0., 2., 0., 2., 2., 2., 4., 2., 2.,$$
$$1., 1., 3., 2., 3., 3., 0., 2., 3., 2., 0., 0., 0., 0., 1., 1, 2., 2., 0., 3.,$$
$$2., 2., 2., 0., 3., 2., 3., 2., 1., 0., 2., 0., 3., 0., 1., 1., 1., 1., 1., 0.\}$$

Grupo	Código
Grupo 1	{1,0,0,0,0}
Grupo 2	{0,1,0,0,0}
Grupo 3	{0,0,1,0,0}
Grupo 4	{0,0,0,1,0}
Grupo 5	{0,0,0,0,1}

A continuación se define los distintos grupos de clasificación, en este caso son 5.

Donde corresponde:

Vector de Salida:

$$Y = \{\{0,0,0,0,1\},\{0,0,0,1,0\},\{0,0,0,0,1\},\{1,0,0,0,0\},\{1,0,0,0,0\},$$
$$\{1,0,0,0,0\},\{1,0,0,0,0\},\{1,0,0,0,0\},\{1,0,0,0,0\},\{1,0,0,0,0\},\{0,1,0,0,0\},$$
$$\{0,1,0,0,0\},\{1,0,0,0,0\},\{0,0,0,0,1\},\{1,0,0,0,0\},\{0,0,0,0,1\},\{0,1,0,0,0\}\}$$

Anexo B

Datos objetivos

Tabla B.1: Tabla 1, con el número de estudiantes de la FCE de la UPO

PROVINCIA		MUNICIPIO		C. POSTAL	
				2765	1
ALMERÍA	3	ALMERÍA CAPITAL	2	4005	1
				4006	1
		BERJA	1	4760	1
BADAJOZ	23	BADAJOZ CAPITAL	7	6001	1
				6004	1
				6006	1
				6011	4
		ALMENDRALEJO	1	6200	1
		VILLALBA DE LOS BARROS	1	6208	1
		VILLAFRANCA DE LOS BARROS	1	6220	1
		FUENTE DE CANTOS	2	6240	2
		MONTEMOLÍN	1	6291	1
		ZAFRA	2	6300	2
		JEREZ DE LOS CABALLEROS	1	6380	1
		DON BENITO	1	6400	1
		MÉRIDA	2	6800	2
		LLERENA	2	6900	2
		PUEBLA DEL MAESTRE	1	6906	1
		BERLANGA	1	6930	1
PALMA	1	PALMA	1	7009	1
CÁCERES	4	CÁCERES CAPITAL	3	10001	2
				10005	1
		PLASENCIA	1	10600	1
CÁDIZ	141	CÁDIZ CAPITAL	14	11001	1
				11003	1
				11004	1
				11007	3
				11008	2

Tabla B.2: Tabla 1 (parte 2)

PROVINCIA		MUNICIPIO		C. POSTAL	
CÁDIZ	141	CÁDIZ CAPITAL	14	11009	4
				11010	1
				11012	1
		SAN FERNANDO	9	11100	9
		CHICLANA DE LA FRONTERA	13	11130	13
		CONIL DE LA FRONTERA	1	11149	1
		VEJER DE LA FRONTERA	1	11150	1
		BARBATE	2	11160	2
		MEDINA SIDONIA	1	11170	1
		BENALUP DE SIDONIA	2	11190	2
		ALGECIRAS	19	11201	4
				11202	1
				11203	2
				11204	1
				11205	3
				11206	2
				11207	6
		LA LÍNEA DE LA CONCEPCIÓN	3	11300	2
				11315	1
		JIMENA DE LA FRONTERA	1	11330	1
		SAN ROQUE	1	11360	1
		PALMORES (TREBUJENA)	2	11379	2
		EL BUZO	2	11390	1
		JEREZ DE LA FRONTERA	32	11401	1
				11402	2
				11403	1
				11405	15
				11406	3
				11407	9
				11408	1

Tabla B.3: Tabla 1 (parte 3)

PROVINCIA		MUNICIPIO		C. POSTAL	
CÁDIZ	141	EL PUERTO DE SANTA MARÍA	10	11500	8
				11530	2
		PUERTO REAL	3	11510	3
		ROTA	5	11520	5
		SANLÚCAR DE BARRAMEDA	10	11540	10
		LA BARCA	1	11570	1
		SAN JOSÉ DEL VALLE	2	11580	2
		UBRIQUE	1	11600	1
		ARCOS DE LA FRONTERA	1	11630	1
		BORNOS	2	11640	1
				11649	1
		VILLAMARTÍN	1	11650	1
		OLVERA	2	11690	2
CIUDAD REAL	1	CIUDAD REAL	1	13200	1
CÓRDOBA	22	CÓRDOBA CAPITAL	6	14002	1
				14003	1
				14005	1
				14008	1
				14011	1
				14012	1
		PEÑARROYA-PUEBLONUEVO	1	14200	1
		AÑORA	1	14450	1
		PUENTE GENIL	1	14500	1
		LA RAMBLA	1	14540	1
		MONTILLA	5	14550	5
		PALMA DEL RÍO	4	14700	4
		HORNACHUELOS	1	14740	1
		AGUILAR DE LA FRONTERA	1	14920	1
		CABRA	1	14940	1

Tabla B.4: Tabla 1 (parte 4)

PROVINCIA		MUNICIPIO		C. POSTAL	
HUELVA	46	HUELVA CAPITAL	15	21001	2
				21002	4
				21003	6
				21004	2
				21005	1
		PUNTA UMBRÍA	1	21100	1
		ALJARAQUE	5	21110	4
				21122	1
		ARACENA	3	21200	3
		LINARES DE LA SIERRA	1	21207	1
		JABUGO	1	21360	1
		ENCINASOLA	1	21390	1
		AYAMONTE	1	21400	1
		ISLA CRISTINA	1	21410	1
		LEPE	2	21440	2
		CARTAYA	1	21450	1
		SAN BARTOLOMÉ DE LA TORRE	1	21510	1
		VALVERDE DEL CAMINO	4	21600	4
		TRIGUEROS	1	21620	1
		BEAS	2	21630	2
		NERVA	1	21670	1
		LA PALMA DEL CONDADO	1	21700	1
		ALMONTE	2	21730	2
		LUCENA DEL PUERTO	1	21820	1
		VILLALBA DEL ALCOR	1	21860	1
JAÉN	2	LA PUERTA DE SEGURA	1	23360	1
		MARMOLEJO	1	23770	1
MÁLAGA	7	MÁLAGA	1	29016	1
		ANTEQUERA	2	29200	2

Tabla B.5: Tabla 1 (parte 5)

PROVINCIA		MUNICIPIO		C. POSTAL	
MÁLAGA	7	RONDA	2	29400	2
		ESTEPONA	1	29680	1
		TORRE DEL MAR	1	29740	1
NAVARRA	1	FIGAROL	1	31311	1
LAS PALMAS	1	ARRECIFE	1	335500	1
VIGO	1	VIGO	1	36207	1
TENERIFE	2	STA. CRUZ	1	38001	1
				38320	1
SEVILLA	1150	SEVILLA CAPITAL	565	41001	13
				41002	5
				41003	23
				41004	15
				41005	18
				41006	35
				41007	21
				41008	18
				41009	15
				41010	43
				41011	60
				41012	41
				41013	104
				41014	7
				41015	14
				41016	7
				41018	53
				41019	4
				41020	69
		MONTEQUINTO	73	41089	73
		CORIA DEL RÍO	14	41100	14

Tabla B.6: Tabla 1 (parte 6)

PROVINCIA		MUNICIPIO		C. POSTAL	
SEVILLA	1150	BOLLULLOS DE LA MITACIÓN	2	41110	2
		ALMENSILLA	1	41111	1
		GELVES	7	41120	7
		LA PUEBLA DEL RÍO	2	41130	2
		ISLA MAYOR	1	41140	1
		ALCALÁ DEL RÍO	4	41200	4
		GUILLENA	5	41210	4
				41219	1
		BURGUILLOS	2	41220	2
		CASTILBLANCO DE LOS ARROYOS	1	41230	1
		ALMADÉN DE LA PLATA	1	41240	1
		LA RINCONADA	3	41300	2
				41309	1
		BRENES	1	41310	1
		CANTILLANA	2	41320	2
		EL PEDROSO	1	41360	1
		CAZALLA DE LA SIERRA	2	41370	2
		GUADALCANAL	1	41390	1
		ÉCIJA	10	41400	10
		CARMONA	4	41410	4
		LA CAMPANA	1	41429	1
		CAÑADA DEL ROSAL	4	41439	4
		LORA DEL RÍO	1	41440	1
		ALCOLEA DEL RÍO	1	41449	1
		PEÑAFLOR	1	41470	1
		ALCALÁ DE GUADAÍRA	57	41500	57
		MAIRENA DEL ALCOR	6	41510	6
		EL VISO DEL ALCOR	3	41520	3

Tabla B.7: Tabla 1 (parte 7)

PROVINCIA		MUNICIPIO		C. POSTAL	
SEVILLA	1150	MORÓN DE LA FRONTERA	5	41530	5
		ESTEPA	2	41560	2
		CASARICHE	1	41580	1
		RODA DE ANDALUCÍA	1	41590	1
		ARAHAL	5	41600	5
		PARADAS	1	41610	1
		MARCHENA	5	41620	5
		OSUNA	3	41640	3
		EL SAUCEJO	1	41650	1
		LOS CORRALES	2	41657	2
		MARTÍN DE LA JARA	1	41658	1
		DOS HERMANAS	72	41700	32
				41701	15
				41702	4
				41703	18
				41704	3
		UTRERA	29	41710	29
		LOS PALACIOS Y VILLAFRANCA	18	41720	17
				41727	1
		LAS CABEZAS DE SAN JUAN	3	41730	1
		LEBRIJA	3	41740	3
		EL CORONIL	3	41760	3
		MONTELLANO	3	41770	3
		SANLÚCAR LA MAYOR	2	41800	2
		OLIVARES	7	41804	7
		UMBRETE	5	41806	5
		ESPARTINAS	15	41807	15
		VILLANUEVA DEL ARISCAL	2	41808	2
		PILAS	8	41840	8

Tabla B.8: Tabla 1 (parte 8)

PROVINCIA		MUNICIPIO		C. POSTAL	
SEVILLA	1150	AZNALCÁZAR	1	41849	1
		VILLAMANRIQUE	1	41850	1
		CAMAS	9	41900	9
		VALENCIANA DE LA CONCEPCIÓN	6	41907	6
		CASTILLEJA DE GUZMÁN	2	41908	2
		SALTERAS	3	41909	3
		SAN JUAN DE AZNALFARACHE	8	41920	8
		MAIRENA DEL ALJARAFE	43	41927	43
		PALOMARES DEL RÍO	9	41928	9
		BORMUJOS	20	41930	20
		TOMARES	49	41940	49
		CASTILLEJA DE LA CUESTA	10	41950	10
		GINES	9	41960	9
		SANTIPONCE	4	41970	4
		LA ALGABA	3	41980	3
TOLEDO	1	MORA	1	45400	1

Fuente: elaboración propia

Anexo C

Resultados parciales

G_k	n_k	mín	máx	l_k	ECM	ϑ_k	r_k	r_{1k}	r_{2k}	ξ
G_1	63	5,2981	9,4231	2	2,3490	2,4495	12	6	6	1,0000
G_2	21	5,5912	7,7618	3	2,2128	2,2361	10	5	5	1,0000
G_3	77	5,0375	9,3350	2	1,9295	2,0000	8	4	4	1,0000
G_4	81	4,9611	9,1472	2	1,6884	1,7321	6	3	3	1,0000
G_5	141	5,0000	9,2350	2	1,0840	1,4142	4	2	2	1,0000
G_6	110	4,7008	9,5000	2	0,6412	1,0000	2	1	1	1,0000
G_7	57	5,2534	8,4196	3	2,1852	2,3452	11	5	6	0,8333
G_8	46	4,7775	8,4792	2	2,0984	2,1213	9	4	5	0,8333
G_9	45	5,0000	7,3833	2	1,8252	1,8708	7	3	4	0,8333
G_{10}	54	5,0000	8,6500	2	1,5330	1,5811	5	2	3	0,8333
G_{11}	37	5,0000	7,5000	2	1,0950	1,2247	3	1	2	0,8333
G_{12}	193	0,0000	0,0000	1	0,0000	0,0000	1	0	1	0,8333
G_{13}	48	5,2167	7,7521	3	2,0829	2,2361	10	4	6	0,6667
G_{14}	32	4,9486	8,4500	3	1,9454	2,0000	8	3	5	0,6667
G_{15}	30	4,9136	6,2366	2	1,6070	1,7321	6	2	4	0,6667
G_{16}	49	5,0000	8,0000	3	1,2436	1,4142	4	1	3	0,6667
G_{17}	94	0,0000	0,0000	1	0,0000	0,0000	2	0	2	0,6667
G_{18}	17	5,0424	8,0736	2	1,9370	2,1213	9	3	6	0,5000
G_{19}	20	5,0048	7,2000	3	1,7839	1,8708	7	2	5	0,5000
G_{20}	29	5,0000	7,4400	4	1,7586	1,5811	5	1	4	0,5000
G_{21}	97	0,0000	0,0000	1	0,0000	0,0000	3	0	3	0,5000
G_{22}	4	5,3500	6,2000	2	1,1566	2,0000	8	2	6	0,3333
G_{23}	6	5,0000	7,0000	2	1,5008	1,7321	6	1	5	0,3333
G_{24}	11	0,0000	0,0000	1	0,0000	0,0000	4	0	4	0,3333
G_{25}	2	5,0000	5,3000	1	0,0000	1,8708	7	1	6	0,1667
G_{26}	1	0,0000	0,0000	1	0,0000	0,0000	5	0	5	0,1667
G_{27}	0	0,0000	0,0000	1	0,0000	0,0000	6	0	6	0,0000
G_{28}	76	0,0000	0,0000	1	0,0000	0,0000	0	0	0	1,0000

S_{kl}	$\overline{S_{kl}}$	n_{kl}	δ	S_{kl}	$\overline{S_{kl}}$	n_{kl}	δ
S_{11}	7,816524306	20	7,81652431	S_{141}	6,805330000	5	4,53688667
S_{12}	6,359999160	43	6,35999916	S_{142}	5,473205247	9	3,64880350
S_{21}	7,246756667	7	7,24675667	S_{143}	5,452702778	18	3,63513519
S_{22}	6,140582222	9	6,14058222	S_{151}	5,611359559	17	3,74090637
S_{23}	5,967499333	5	5,96749933	S_{152}	5,291861538	13	3,52790769
S_{31}	7,834751003	27	7,83475100	S_{161}	5,688839286	28	3,79255952
S_{32}	5,979807625	50	5,97980763	S_{162}	5,531333333	10	3,68755556
S_{41}	6,733698830	38	6,73369883	S_{163}	5,357545455	11	3,57169697
S_{42}	5,615837532	43	5,61583753	S_{171}	0,00000000	94	0,66666667
S_{51}	7,755457083	60	7,75545708	S_{181}	6,086750000	11	3,04337500
S_{52}	5,915775720	81	5,91577572	S_{182}	5,529706019	6	2,76485301
S_{61}	7,520357143	28	7,52035714	S_{191}	5,805928571	7	2,90296429
S_{62}	5,515869919	82	5,51586992	S_{192}	5,601982955	11	2,80099148
S_{71}	6,722107867	25	5,60175656	S_{193}	5,023125000	2	2,51156250
S_{72}	6,058116667	11	5,04843056	S_{201}	5,794023810	7	2,89701190
S_{73}	5,671280159	21	4,72606680	S_{202}	5,318181818	11	2,65909091
S_{81}	6,878001705	22	5,73166809	S_{203}	5,307416667	6	2,65370833
S_{82}	5,753862066	24	4,79488505	S_{204}	5,176000000	5	2,58800000
S_{91}	6,618311508	14	5,51525959	S_{211}	0,00000000	97	0,50000000
S_{92}	5,631392921	31	4,69282743	S_{221}	5,850000000	2	1,95000000
S_{101}	5,917606349	42	4,93133862	S_{222}	5,512500000	2	1,83750000
S_{102}	5,663923611	12	4,71993634	S_{231}	6,450000000	4	2,15000000
S_{111}	5,750055172	29	4,79171264	S_{232}	5,125000000	2	1,70833333
S_{112}	5,550000000	8	4,62500000	S_{241}	0,00000000	11	0,33333333
S_{121}	0,000000000	193	0,83333333	S_{251}	5,150000000	2	0,85833333
S_{131}	5,970130903	30	3,98008727	S_{261}	0,00000000	1	0,16666667
S_{132}	5,830167535	12	3,88677836	S_{271}	0,00000000	0	0,00000000
S_{133}	5,424850694	6	3,61656713	S_{281}	0,00000000	76	0,00000000

ξ_k	S_{kl}	$\overline{S_{kl}}$	ξ_k	S_{kl}	$\overline{S_{kl}}$
1	S_{31}	7,834751003	0,8333	S_{102}	5,663923611
1	S_{11}	7,816524306	0,8333	S_{92}	5,631392921
1	S_{51}	7,755457083	1	S_{42}	5,615837532
1	S_{61}	7,520357143	0,6667	S_{151}	5,611359559
1	S_{21}	7,246756667	0,5	S_{192}	5,601982955
0,8333	S_{81}	6,878001705	0,8333	S_{112}	5,550000000
0,6667	S_{141}	6,805330000	0,6667	S_{162}	5,531333333
1	S_{41}	6,733698830	0,5	S_{182}	5,529706019
0,8333	S_{71}	6,722107867	1	S_{62}	5,515869919
0,8333	S_{91}	6,618311508	0,3333	S_{222}	5,512500000
0,3333	S_{231}	6,450000000	0,6667	S_{142}	5,473205247
1	S_{12}	6,359999160	0,6667	S_{143}	5,452702778
1	S_{22}	6,140582222	0,6667	S_{133}	5,424850694
0,5	S_{181}	6,086750000	0,6667	S_{163}	5,357545455
0,8333	S_{72}	6,058116667	0,5	S_{202}	5,318181818
1	S_{32}	5,979807625	0,5	S_{203}	5,307416667
0,6667	S_{131}	5,970130903	0,6667	S_{152}	5,291861538
1	S_{23}	5,967499333	0,5	S_{204}	5,176000000
0,8333	S_{101}	5,917606349	0,1667	S_{251}	5,150000000
1	S_{52}	5,915775720	0,3333	S_{232}	5,125000000
0,3333	S_{221}	5,850000000	0,5	S_{193}	5,023125000
0,6667	S_{132}	5,830167535	0,8333	S_{121}	0,000000000
0,5	S_{191}	5,805928571	0,3333	S_{241}	0,000000000
0,5	S_{201}	5,794023810	0,1667	S_{261}	0,000000000
0,8333	S_{82}	5,753862066	0	S_{271}	0,000000000
0,8333	S_{111}	5,750055172	0	S_{281}	0,000000000
0,6667	S_{161}	5,688839286	0,6667	S_{171}	0,000000000
0,8333	S_{73}	5,671280159	0,5	S_{211}	0,000000000

ξ_k	S_{kl}	δ	ξ_k	S_{kl}	δ
1	S_{31}	7,83475100	0,6667	S_{151}	3,74090637
1	S_{11}	7,81652431	0,6667	S_{162}	3,68755556
1	S_{51}	7,75545708	0,6667	S_{142}	3,64880350
1	S_{61}	7,52035714	0,6667	S_{143}	3,63513519
1	S_{21}	7,24675667	0,6667	S_{133}	3,61656713
1	S_{41}	6,73369883	0,6667	S_{163}	3,57169697
1	S_{12}	6,35999916	0,6667	S_{152}	3,52790769
1	S_{22}	6,14058222	0,5	S_{181}	3,04337500
1	S_{32}	5,97980763	0,5	S_{191}	2,90296429
1	S_{23}	5,96749933	0,5	S_{201}	2,89701190
1	S_{52}	5,91577572	0,5	S_{192}	2,80099148
0,8333	S_{81}	5,73166809	0,5	S_{182}	2,76485301
1	S_{42}	5,61583753	0,5	S_{202}	2,65909091
0,8333	S_{71}	5,60175656	0,5	S_{203}	2,65370833
1	S_{62}	5,51586992	0,5	S_{204}	2,58800000
0,8333	S_{91}	5,51525959	0,5	S_{193}	2,51156250
0,8333	S_{72}	5,04843056	0,3333	S_{231}	2,15000000
0,8333	S_{101}	4,93133862	0,3333	S_{221}	1,95000000
0,8333	S_{82}	4,79488505	0,3333	S_{222}	1,83750000
0,8333	S_{111}	4,79171264	0,3333	S_{232}	1,70833333
0,8333	S_{73}	4,72606680	0,1667	S_{251}	0,85833333
0,8333	S_{102}	4,71993634	0,8333	S_{121}	0,83333333
0,8333	S_{92}	4,69282743	0,6667	S_{171}	0,66666667
0,8333	S_{112}	4,62500000	0,5	S_{211}	0,50000000
0,6667	S_{141}	4,53688667	0,3333	S_{241}	0,33333333
0,6667	S_{131}	3,98008727	0,1667	S_{261}	0,16666667
0,6667	S_{132}	3,88677836	0	S_{271}	0,00000000
0,6667	S_{161}	3,79255952	0	S_{281}	0,00000000

Índice de figuras

2.1. Neurona biológica, modificada de http://es.123rf.com/ 17

2.2. RNA monocapa 48

2.3. RNA multicapa 49

2.4. RNA recurrente 50

2.5. Asociador lineal 54

2.6. Perceptrón simple 55

2.7. RNA Adaline 56

2.8. RNA Madaline 58

2.9. Función de base radial 60

2.10. Red de Hopfield 61

2.11. Red probabilística 64

2.12. Mapa autoorganizado 67

2.13. Máquina de Boltzmann 68

2.14. RNA adaptive resonance theory network 69

2.15. RNA de memoria asociativa bidireccional 70

2.16. Counter Propagation Networks 72

3.1. Ilustración de una clasificación mediante RNA por subgrupos . . . 95

3.2. Ilustración de un ejemplo de clasificación mediante RNA por subgrupos 99

3.3. Pantalla de inicio del programa propuesto en Mathematica 9 104

3.4. Definición de parámetros 112

3.5. Ejemplo de RNA adecuada, seleccionada por el programa informático 136

3.6. Ejemplo de la RNA utilizada, con el conjunto completo de conexiones iniciales 141

3.7. Ejemplo de la RNA utilizada tras la simplificación 142

5.1. Provincias españolas con algún alumno matriculado por primera vez en la FCE de la UPO durante los años 2009-2012 192

5.2. Provincias andaluzas según los alumnos matriculados por primera vez en la FCE de la UPO durante los años 2009-2012 194

5.3. Municipios sevillanos según los alumnos matriculados por primera vez en la FCE de la UPO durante los años 2009-2012 197

5.4. Relación entre la población de los municipios de la provincia de Sevilla y los estudiantes de la FCE de la UPO 200

5.5. Relación entre la población de los municipios de la provincia de Sevilla (sin la capital) y los estudiantes de la FCE de la UPO . . . 201

5.6. Relación entre la población de Sevilla capital y los municipios de la Corona 1 y los estudiantes de la FCE de la UPO 203

5.7. Relación entre la población de los distritos de Sevilla capital y el número de estudiantes de la FCE de la UPO 204

5.8. Relación entre la distancia a la Universidad y los estudiantes de la FCE de la UPO 206

5.9. Relación entre la distancia a la Universidad y los estudiantes de la FCE de la UPO provincia de Sevilla sin la Capital 207

5.10. Relación entre la distancia a la Universidad y los estudiantes de la FCE de la UPO de la Corona 1 208

5.11. Relación entre la distancia a la Universidad y los estudiantes de la FCE de la UPO por distritos 209

5.12. Relación entre la distancia a la Universidad y los estudiantes de la FCE de la UPO por código postal 210

5.13. Relación entre el tiempo a la Universidad y los estudiantes de la FCE de la UPO de Sevilla 211

5.14. Relación entre el tiempo a la Universidad y los estudiantes de la FCE de la UPO de Sevilla Capital 212

5.15. Relación entre el tiempo a la Universidad y los estudiantes de la FCE de la UPO de Sevilla Capital y la Corona 1 213

5.16. Relación entre el tiempo a la Universidad y los estudiantes de la FCE de la UPO de Sevilla por distritos 214

5.17. Relación entre el tiempo a la Universidad y los estudiantes de la FCE de la UPO de Sevilla por código postal 214

5.18. Mapa de Sevilla con las paradas del metro - línea 1 215

5.19. Mapa de Sevilla por códigos postales 216

5.20. Mapa de Sevilla según el porcentaje de estudiantes de la FCE de la UPO 216

Índice de tablas

2.1. Relación entre el cerebro humano y el ordenador 19

3.1. Resultado de entrenar una RNA con neuronas en la capa oculta . . 101

3.2. Definición de los parámetros para buscar la RNA más adecuada . . 121

3.3. Resultados de los entrenamientos de las RNA del ejemplo de búsqueda 124

3.4. Resultados de los entrenamientos de las RNA del ejemplo de búsqueda (parte 2) 125

3.5. Resultados de los entrenamientos de las RNA del ejemplo de búsqueda (parte 3) 126

3.6. Resultados de los entrenamientos de las RNA del ejemplo de búsqueda (parte 4) 127

3.7. Resultados de los entrenamientos de las RNA del ejemplo de búsqueda (parte 5) 128

3.8. Resultados de los entrenamientos de las RNA del ejemplo de búsqueda (parte 6) 129

3.9. Resultados de los entrenamientos de las RNA del ejemplo de búsqueda (parte 7) 130

3.10. Resultados de los entrenamientos de las RNA del ejemplo de búsqueda (parte 8) 131

3.11. Mejores RNA del ejemplo de búsqueda, ordenadas por el logaritmo del error de validación 132

3.12. Mejores RNA del ejemplo de búsqueda, ordenadas según el porcentaje de error de validación 133

5.1. Asignaturas de formación básica u obligatoria impartidas por el Área de Métodos Cuantitativos en GADE 161

5.2. Asignaturas de formación básica u obligatoria impartidas por el Área de Métodos Cuantitativos en GADE-GD 162

5.3. Asignaturas de formación básica u obligatoria impartidas por el Área de Métodos Cuantitativos en GFC 162

5.4. Asignaturas de formación básica u obligatoria impartidas por el Área de Métodos Cuantitativos en GFC-GD 163

5.5. Tabla de frecuencia por titulaciones 187

5.6. Tabla de frecuencia por curso académico 188

5.7. Tabla de frecuencia por sexo 188

5.8. Tabla de frecuencia por edad 189

5.9. Tabla de frecuencia por tipo de centro 189

5.10. Tabla de frecuencia por tipo de acceso a la Universidad 190

5.11. Provincias de las que procede algún alumno (según el año) 193

5.12. Número de alumnos matriculados por primera vez en la FCE de la UPO en 2009-2012 por provincias andaluzas 195

5.13. Número de alumnos matriculados por primera vez en la FCE de la UPO en 2009-2012 por municipios sevillanos 196

5.14. Número de alumnos matriculados por primera vez en la FCE de la UPO en 2009-2012 por municipios de la Corona 1 198

5.15. Distribución porcentual del nº de estudiantes y la población de los municipios de la Corona 1 y los distritos de Sevilla capital 202

5.16. Distribución porcentual del nº de estudiante y la población de los municipios de la Corona 1 y los distritos de Sevilla capital 205

5.17. Distribución porcentual del nº de estudiantes según la existencia de metro y por código postal 215

6.1. Clasificación de los subgrupos y valor del índice 239

6.2. Correlación entre variables predictivas, media aritmética e índice . 239

6.3. Porcentaje de pronósticos incorrectos para cada uno de los grupos (establecidos según δ) 240

6.4. Resultado del entrenamiento de una RNA con dos capas ocultas (para estimar el rendimiento académico por titulación) 240

6.5. Variables recodificadas para fsQCA 250

B.1. Tabla 1, con el número de estudiantes de la FCE de la UPO 285

B.2. Tabla 1 (parte 2) 286

B.3. Tabla 1 (parte 3) 287

B.4. Tabla 1 (parte 4) 288

B.5. Tabla 1 (parte 5) 289

B.6. Tabla 1 (parte 6) 290

B.7. Tabla 1 (parte 7) 291

B.8. Tabla 1 (parte 8) 292